# Data Science for the Health Sciences

# Data Science for the Health Sciences

Foundations, Statistical Inference, Machine Learning, and Infectious Disease Modelling with Applications to Nigerian Health Systems with Worked Examples, Mini-Projects, and Answers to All End-of-Chapter Exercises

**Oluwafemi Samuel Oyamakin**

Stirling-Horden Publishers Ltd.
No 6 Adonai Estate, Opposite
Crown Height College,
Alapata, Igboloyin Road,
Ojoo, Ibadan.
Phone: 08023519154
stirlinghorden@ymail.com; info@stirlinghorden.com
www.stirlinghorden.com

University Post Office
Box 20984
Ibadan
Oyo State, Nigeria

7 Ikioda Street
Jattu-Uzairue
Edo State, Nigeria

© Oluwafemi Samuel Oyamakin

First Edition: 2026

ISBN: 978-978-035

*All Rights Reserved*

Apart from any fair dealing for the purpose of research or private study, criticism or review, as permitted under the Copyright, Designs and Patents Act of Nigeria, this publication may not be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licencing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

Stirling-Horden Publishers Ltd.
Lagos, Ibadan, Benin City, Jattu-Uzairue

# Dedication

# Acknowledgements

This first edition benefited from the sustained encouragement of colleagues across the Department of Statistics, University of Ibadan, and from collaborators on the NiLi Platform, the Nigeria Malaria Modelling Fellowship, and the several institutional cohort studies whose data-generating realities shaped the worked examples throughout this book. Any remaining errors are the author's alone.

# Preface

This textbook grew out of a simple, recurring frustration: excellent data science textbooks exist in abundance, and excellent epidemiology and biostatistics textbooks exist in abundance, but very little bridges the two using the specific data realities of the Nigerian health system — the strengths and quirks of DHIS2/NHMIS reporting, the design of the Nigeria Demographic and Health Survey, the diagnostic surveillance architecture of the National Malaria Elimination Programme, and the practical challenges of running a respondent-driven sampling study or a longitudinal adolescent health cohort in a Nigerian setting.

This book is written for three overlapping audiences: postgraduate students and researchers in Statistics, Epidemiology, Public Health, and related fields who need a single, connected treatment running from probability foundations through machine learning and infectious disease modelling; working monitoring-and-evaluation and health informatics professionals who handle DHIS2 and survey data daily and wish to extend their practice toward rigorous inference and predictive modelling; and computer scientists and data engineers entering the health sector who need a grounded introduction to epidemiological reasoning and research ethics alongside their existing technical skills.

The structure follows a single pipeline — introduced in Chapter 1 and returned to throughout — from problem formulation, through data collection and cleaning, exploratory analysis, statistical inference, regression and survival modelling, time series forecasting, machine learning, unsupervised learning, mechanistic infectious disease modelling, spatial epidemiology, and modern big-data/AI methods, closing with a chapter on data

ethics and governance and a capstone chapter of worked case studies. Code examples are given in both R and Python, reflecting the reality that a working Nigerian health data scientist encounters both in the field.

It is my hope that this book equips its readers not merely to run a procedure, but to ask the harder and more durable question that this book returns to in every chapter: where did this data come from, and what does that mean for how much I should trust the number in front of me.

**How to Use This Book**

This first edition adds, to every chapter, one or more fully worked numerical examples, additional code demonstrations in both R and Python, a hands-on mini-project, and an expanded set of end-of-chapter exercises, together with a complete appendix of model answers to every exercise in the book. A reader working through this book for the first time is encouraged to read a chapter's conceptual sections in full, attempt the worked examples independently before reading the book's own working, and complete the mini-project as the most substantial and realistic test of whether the chapter's material has genuinely been absorbed.

A reader using this book as a reference, rather than reading cover to cover, will find that each chapter's summary box and mini-project can be read on their own to quickly assess whether a given chapter addresses the problem at hand, before committing time to the full chapter text. The answers appendix at the back of the book is organised by chapter and exercise number for quick lookup, and it is deliberately written to show the reasoning behind each answer, not just a final numeric result, so that it remains useful even to a reader who has not attempted the corresponding exercise first.

Instructors adopting this book for a semester-length course will find that the eighteen chapters map naturally onto an eighteen-week teaching schedule, with each chapter's mini-project suitable as a weekly practical assignment and the capstone chapter (17) and Python chapter (18) suitable as the basis for a final project spanning the last two to three weeks of a course.

This edition also reflects a deliberate editorial choice worth stating plainly: every worked example, case vignette, and mini-project in this book uses simulated or illustrative data constructed to reflect the real, well-documented statistical patterns found in Nigerian health systems (such as the HMIS-versus-register test positivity discrepancy discussed throughout the early chapters), rather than reproducing any specific individual's or institution's actual confidential dataset. Readers applying this book's methods to their own real data should expect the same qualitative lessons to hold, while the exact numbers in any given worked example should be understood as pedagogically constructed rather than drawn from a specific published study.

**Olúwáfémi Samuel Oyámákin, PhD, CStat., CSci.**
*Department of Statistics, University of Ibadan*

# Notation and Abbreviations

This reference page collects the abbreviations and notational conventions used repeatedly across the book's eighteen chapters, for quick lookup without needing to locate the chapter in which a term was first introduced.

| | |
|---|---|
| **ANC** | Antenatal Care |
| **ART** | Antiretroviral Therapy |
| **AUC** | Area Under the (ROC) Curve |
| **BYM** | Besag-York-Mollié (spatial smoothing model) |
| **CFR** | Case Fatality Ratio |
| **CI** | Confidence Interval (or Credible Interval, in a Bayesian context) |
| **CNN** | Convolutional Neural Network |
| **DHIS2** | District Health Information Software, version 2 |
| **EDA** | Exploratory Data Analysis |
| **FCT** | Federal Capital Territory |
| **FMoH** | Federal Ministry of Health |
| **GLM** | Generalised Linear Model |
| **HMIS** | Health Management Information System |
| **IDSR** | Integrated Disease Surveillance and Response |
| **LGA** | Local Government Area |
| **MAR / MCAR / MNAR** | Missing at Random / Missing Completely at Random / Missing Not at Random |
| **MICS** | Multiple Indicator Cluster Survey |
| **M&E** | Monitoring and Evaluation |

| | |
|---|---|
| **NCDC** | Nigeria Centre for Disease Control |
| **NDHS** | Nigeria Demographic and Health Survey |
| **NLP** | Natural Language Processing |
| **NTD** | Neglected Tropical Disease |
| **PCA** | Principal Component Analysis |
| **PPV/NPV** | Positive / Negative Predictive Value |
| **RDS** | Respondent-Driven Sampling |
| **RDT** | Rapid Diagnostic Test |
| **R0/Rt** | Basic / Effective Reproduction Number |
| **SEIR/SVEIHR** | Susceptible-Exposed-Infectious-Recovered (model), with Vaccinated/ Hospitalised extensions |
| **SORMAS** | Surveillance Outbreak Response Management and Analysis System |
| **TPR** | Test Positivity Rate |

# Contents

**Dedication** **v**

**Acknowledgements** **vi**

**Preface** **vii**

**Notation and Abbreviation** **x**

**PART I: Foundations**

**Chapter 1:** Introduction to Data Science for the Health Sciences **3**

**Chapter 2:** The Nigerian Health Data Ecosystem **21**

**Chapter 3:** Foundations of Probability and Statistical Thinking for Health Data **37**

**PART II: Data Wrangling and Exploratory Analysis**

**Chapter 4:** Data Management: Types, Sources, Collection, and Cleaning **55**

**Chapter 5:** Exploratory Data Analysis and Visualization **71**

**Chapter 6:** Programming Foundations in R and Python for Health Data Science **86**

**PART III: Statistical Inference and Modelling**

**Chapter 7:** Statistical Inference and Hypothesis Testing in Health Research **103**

**Chapter 8:** Regression Modelling: Linear, Logistic, and Beyond **118**

**Chapter 9:** Survival Analysis and Time-to-Event Data **134**

**Chapter 10:** Time Series Analysis and Disease Surveillance Forecasting **149**

**PART IV: Machine Learning and Advanced Methods**

**Chapter 11:** Machine Learning Fundamentals for Health Prediction **166**

**Chapter 12:** Unsupervised Learning: Clustering and Dimensionality Reduction **181**

**Chapter 13:** Mathematical Modelling of Infectious Diseases **195**

**Chapter 14:** Spatial Epidemiology and Disease Mapping **211**

**PART V: Big Data, Ethics, and Practice**

**Chapter 15:** Big Data, AI, and Deep Learning in Health Systems **227**

**Chapter 16:** Data Ethics, Governance, and the Future of Health Data Science in Nigeria **242**

**Chapter 17:** Capstone: Case Studies and a Framework for Independent Practice **256**

**PART VI: Python for Disease Epidemiology**

**Chapter 18:** Introduction to Python for Disease Epidemiology **273**

**Glossary of Key Terms** **293**

**Selected References and Further Reading** **298**

**Appendix:** Answers to End-of-Chapter Exercises (All 18 Chapters) **301**

# Part 1
## Foundations

# 1 Introduction to Data Science for the Health Sciences

## What is Data Science, and Why Does Health Need It?

Data science is the disciplined practice of extracting reliable knowledge and defensible decisions from data. It sits at the intersection of three older disciplines — statistics, computer science, and domain (subject-matter) expertise — and adds to them a fourth ingredient that is often underestimated: the engineering discipline required to move data from a messy, real-world source into a form that can be analysed, modelled, and acted upon reproducibly. In the health sciences, this fourth ingredient is not a luxury. A clinician who trusts a p-value produced by a script that silently dropped half the missing records, or a policymaker who allocates malaria commodities on the basis of a chart built from an unversioned spreadsheet, is exposed to a specific and preventable kind of error: an error not of statistical theory, but of data handling.

This textbook treats data science as a single connected pipeline rather than a collection of unrelated tools. That pipeline runs from problem formulation, through data collection and cleaning, to exploratory analysis, formal statistical inference, predictive modelling (including machine learning), and finally to communication and decision support. Every chapter in this book

will return to that pipeline and show where its topic sits within it, because a technique learned in isolation — a regression coefficient, a clustering algorithm, a neural network — is of little use to a health system unless the analyst also understands where the data came from, what selection processes shaped it, and who will act on the result.

Health data differs from data in many other domains in ways that matter for how we teach and practise data science. It is frequently incomplete by design (a patient who does not return for a scheduled visit is missing not at random). It is subject to strict ethical and legal constraints (informed consent, the National Health Act 2014, and institutional review board approval in the Nigerian context). It mixes structured records (a DHIS2 aggregate indicator, a laboratory result field) with unstructured ones (a clinician's free-text note, a caregiver's verbal description of symptoms). And, critically, the consequences of getting the analysis wrong are not abstract: a miscalibrated risk score can mean a child is not referred for emergency care, and an over-optimistic forecast of vaccine demand can mean a stock-out in a rural facility. Data science for the health sciences must therefore be learned with an unusual level of rigour and humility.

## The Data Science Pipeline: From Question to Decision

A useful mental model for the whole of this book is the pipeline shown below. It is deliberately drawn as a loop rather than a straight line, because in practice health data science is iterative: an unexpected pattern found during exploratory analysis routinely sends the analyst back to the field to ask a better question or collect an additional variable.

- Problem formulation: translating a health system question ("are we seeing more severe malaria cases in the rainy season in Kwara State than the dry season?") into an answerable data question.

- Data collection and provenance: identifying which data source(s) — HMIS/DHIS2 registers, a household survey such as the Nigeria Demographic and Health Survey (NDHS), a cohort dataset, or a purpose-built study instrument — can answer the question, and documenting exactly how each variable was captured.
- Data cleaning and management: detecting and handling missing values, duplicate records, implausible entries (a reported age of 200 years, a temperature of 15 degrees Celsius in a febrile patient), and reconciling identifiers across linked datasets.
- Exploratory data analysis (EDA): summarising, tabulating, and visualising the cleaned data to generate and refine hypotheses before any formal model is fitted.
- Formal analysis and modelling: applying the statistical or machine learning method appropriate to the question, the data structure, and the assumptions that can reasonably be defended.
- Validation and communication: checking model assumptions and predictive performance, and translating results into a form that a Local Government Area (LGA) disease control officer, a hospital administrator, or a Federal Ministry of Health (FMoH) technical working group can act on.
- Decision and feedback: the point at which the analysis changes a policy, a clinical protocol, or a resource allocation — and where new data begins to accumulate that will eventually be fed back into the pipeline.

Throughout this book, worked examples will be anchored to Nigerian health data contexts that recur across the chapters: routine health facility data reported through DHIS2/NHMIS, malaria diagnostic surveillance data from Global Fund-supported

states, neglected tropical disease (NTD) surveillance (in particular cutaneous and visceral leishmaniasis), adolescent health cohort data, respondent-driven samples of key populations, and dermatoglyphic (fingerprint pattern) data collected in partnership with clinical referral centres. These are not hypothetical illustrations; they are the kind of dataset a Nigerian health data scientist will actually be handed.

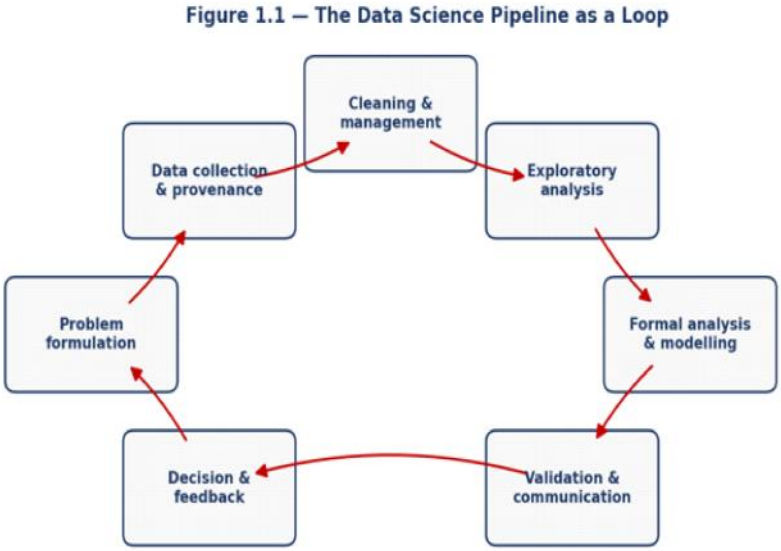


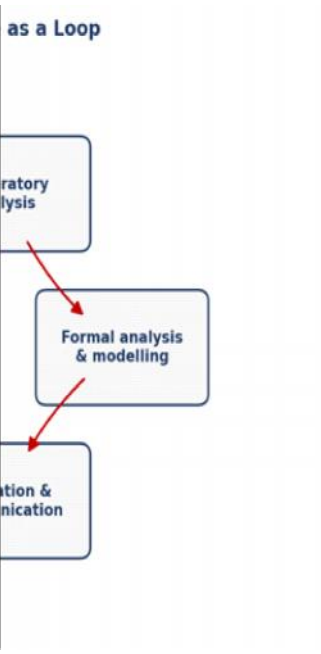

*Fig. 1.1 — The data science pipeline, drawn as a loop (Section 1.2).*

## Who Does Data Science in a Health System, and What Do they Need to Know?

It is useful to distinguish three overlapping roles that this book serves simultaneously, because the balance of statistical theory, programming, and domain framing differs across them.

- The biostatistician or epidemiologist who needs the machine learning and big-data chapters of this book to extend a classical training in inference to modern predictive and unsupervised methods.

- The health informatics or monitoring-and-evaluation (M&E) professional who works daily with DHIS2 and survey data and needs a rigorous statistical foundation to move beyond descriptive dashboards.
- The computer scientist or data engineer entering the health sector who needs a grounding in epidemiological reasoning, study design, and the ethical framework of health research before their technical skills can be trusted with patient-level data.




A reader from any of these three backgrounds should be able to follow this book, because each chapter opens with the conceptual foundation before introducing computational implementation in R and Python, the two languages that dominate modern health data science and that are used interchangeably by working analysts in Nigerian research institutions, universities, and public health agencies.

## Case Vignette

### Why the Pipeline Matters

Consider a real class of problem that recurs in Nigerian malaria control: comparing the test positivity rate (TPR) derived from routine HMIS reporting against the TPR verified from source registers and laboratory logs across the states supported by the Global Fund. On the surface, this looks like a simple two-sample comparison. In practice, a naive analyst who simply merges the two spreadsheets and runs a t-test will be misled, because the two data streams differ systematically in when they aggregate cases (weekly versus monthly), how they treat clinically diagnosed (unconfirmed) cases, and how consistently facilities report in a given month. A rigorous treatment — of the kind developed in Chapters 4, 5, and 8 of this book — reveals a consistent and policy-relevant finding: HMIS-derived TPR tends to run systematically higher than archive-verified TPR across states, a discrepancy with

direct implications for commodity forecasting and for how confidently a State Malaria Elimination Programme can rely on routine data alone. This example illustrates the central argument of this chapter: the statistical technique is often the easiest part of the problem; understanding and correcting for how the data was generated is what makes the analysis trustworthy.

## Worked Example

### Comparing Two 'Equivalent' Numbers that are Not Equivalent

Suppose a facility's monthly HMIS summary report shows 340 malaria tests performed and 210 confirmed positive, for an HMIS-derived TPR of 61.8%. A retrospective audit of the same month's laboratory register at the same facility counts 402 tests performed and 218 confirmed positive, for a register-verified TPR of 54.2%. The two numbers differ by 7.6 percentage points, and the difference is not explained by a data-entry typo: the HMIS summary form, in this example, was completed before the last five days of the month were transcribed into the register, so it systematically under-counted the (mostly negative) tests performed in the final week, inflating the reported positivity rate. This is a textbook illustration of the point made in Section 1.4: the two 'TPR' figures are not measuring the same underlying quantity with different amounts of noise; they are measuring the numerator and denominator of the same rate at two different, non-random points in the reporting cycle. A naive analyst comparing these two numbers across dozens of facilities and concluding that 'HMIS over-reports positivity by roughly 7 percentage points on average' would be correct about the direction and rough size of the bias, but should still trace at least a sample of the discrepancy to its root cause (as done here) before generalising the finding to a policy recommendation.

| Quantity | HMIS summary report | Register audit | Difference |
|---|---|---|---|
| Tests performed | 340 | 402 | -62 (-15.4%) |
| Confirmed positive | 210 | 218 | -8 (-3.7%) |
| Test positivity rate | 61.8% | 54.2% | +7.6 pp |

## A First Look at Code: Reading and Inspecting a Health Dataset

Before any of the statistical machinery introduced later in this book can be applied, an analyst must be able to load a dataset and perform the most basic checks on it: how many rows and columns does it have, what are the variable names and types, and are there any glaringly implausible values? The two code blocks below perform this first inspection in R and in Python on a simulated facility-level malaria dataset of the kind introduced in this chapter, and they will be extended throughout the book as the running example dataset for many end-of-chapter exercises.

***R example***

```
# R: first inspection of a facility-level malaria dataset
malaria <- read.csv("hmis_malaria_monthly.csv")

dim(malaria)                # number of rows and columns
str(malaria)                # variable names and types
summary(malaria$tests_performed)
summary(malaria$confirmed_positive)

# Quick plausibility check: TPR should lie between 0 and 1
malaria$tpr <- malaria$confirmed_positive / malaria$tests_performed
range(malaria$tpr, na.rm = TRUE)
sum(malaria$tpr > 1, na.rm = TRUE)   # count of facility-months with an impossible TPR
```

***PYTHON example***

```
# Python: equivalent first inspection
import pandas as pd

malaria = pd.read_csv("hmis_malaria_monthly.csv")

print(malaria.shape)                # (rows, columns)
print(malaria.dtypes)               # variable types
print(malaria[["tests_performed", "confirmed_positive"]].describe())

malaria["tpr"] = malaria["confirmed_positive"] / malaria["tests_performed"]
print(malaria["tpr"].min(), malaria["tpr"].max())
print((malaria["tpr"] > 1).sum())    # facility-months with an impossible TPR
```

## A Note on this Book's Structure and How to Read It

Each chapter in this book follows a consistent structure that mirrors the pipeline described in Section 1.2: conceptual foundations first, followed by one or more fully worked numerical examples, followed by code demonstrations in R and/or Python, and closing with a chapter summary, a mini-project, and a set of end-of-chapter exercises with model answers collected in the appendix at the back of the book. Readers who are already comfortable with a chapter's foundational material are encouraged to move directly to its worked examples and mini-project, since these are where the book's Nigerian health data framing is most concretely demonstrated and where the most durable learning tends to occur. Readers newer to a topic are encouraged to work through a chapter's exercises before checking the appendix answer, since the productive struggle of attempting an answer first is a well-established aid to retention that mere reading does not provide.

## Extended Case Study

### Two Analysts, One Dataset, Two Different Answers

To make the abstract argument of this chapter concrete, consider a scenario that recurs, in one form or another, in almost every Nigerian health data science team: two analysts are independently given the same raw facility-month malaria dataset and asked to answer the same question — 'did test positivity fall after the state's indoor residual spraying (IRS) campaign?' Analyst A loads the spreadsheet, computes the mean TPR for the three months before and the three months after the campaign's stated start date, finds a fall from 38% to 29%, and reports this as clear evidence of programme impact. Analyst B, working from the same raw file, first checks the number of facilities reporting in each month (finding that reporting completeness rose from 61% to 84% of facilities over the same period, likely for reasons unrelated to

the spraying campaign itself), checks whether the campaign's actual roll-out date varied by LGA rather than occurring uniformly on the stated state-wide date, and re-examines the pre-period for a seasonal decline that was already underway before the campaign began (the pre-period spans the tail end of the rainy season, when transmission is naturally beginning to fall regardless of any intervention).

After this additional work, Analyst B's more careful reconstruction shows that roughly half of the apparent 9-percentage-point fall is attributable to normal seasonal decline that would have occurred regardless of the IRS campaign, and a further portion is confounded by the change in which (generally lower-transmission, better-resourced) facilities were reporting in the later period. The genuine, defensible estimate of the campaign's effect — after adjusting for season and reporting-completeness changes using the regression methods of Chapter 8 — is a more modest but still meaningful 3-4 percentage point reduction. Both analysts used entirely correct arithmetic; the difference between their answers is entirely a matter of data-generating-process discipline of exactly the kind this chapter has argued for. When this dataset was actually used to brief a State Ministry of Health technical working group, it was Analyst B's more cautious, adjusted estimate that was ultimately presented, and it changed the programme's messaging from 'IRS delivered a dramatic 24% relative reduction in positivity' to the more defensible and, in the long run, more credible 'IRS appears to have delivered a modest but real reduction in positivity, on top of an already-expected seasonal decline' — a less dramatic headline, but one that would not need to be walked back under later scrutiny.

## Building a Personal Habit of Provenance-First Analysis

The single most transferable skill this opening chapter can offer is not any specific technique but a habit: before opening a dataset

in R, Python, or Excel, pausing to write down, in one or two sentences, exactly what population and time period the data claims to describe, how it was collected, and who would be harmed by a wrong conclusion drawn from it. This costs perhaps two minutes and yet reliably prevents a large fraction of the most common and most damaging errors in applied health data science — the kind illustrated throughout this book, from the malaria TPR discrepancy of Section 1.5 to the reverse-causation pitfall of Section 8.10. Analysts who adopt this habit early in their careers report that it becomes automatic within a few months, at which point it stops feeling like an extra step and starts feeling like an inseparable part of what 'starting an analysis' actually means.

A simple way to operationalise this habit, useful for individuals and teams alike, is a short standing checklist kept visible during any analysis session: What exactly does each row represent? What time period and geography does this cover? What is the known or suspected direction of any bias in how this was collected? Who will read or act on this analysis, and what would they need to know about its limitations to use it responsibly? Answering these four questions in writing, however briefly, before any code is run, converts an implicit and easily-skipped judgment call into an explicit, reviewable step — precisely the discipline that separates the confident but potentially misleading Analyst A of Section 1.8 from the more careful and ultimately more useful Analyst B.

## The Cost of Analysis Paralysis

### Balancing Rigour and Timeliness

Everything argued for in this chapter — provenance checks, careful data-generating-process scrutiny, resisting premature conclusions — could, taken to an extreme, produce an analyst who never delivers a timely answer because every dataset can always be interrogated further. A State Ministry of Health facing an active outbreak genuinely cannot wait three weeks for a

perfectly validated, triangulated, publication-grade analysis before deciding whether to deploy emergency supplies; a real-time decision often has to be made on the best currently available evidence, explicitly flagged with its limitations, rather than withheld until an ideal analysis is complete. The skill this book aims to build is not an unconditional demand for perfect data before any conclusion is drawn, but a calibrated judgment: understanding which limitations are severe enough to invalidate a conclusion outright, which are manageable with an explicit caveat, and which can reasonably be set aside given the time pressure and stakes of a specific decision — a judgment that itself improves with experience and that this book's worked examples are intended to help develop.

## Building a Personal Portfolio from this Book's Mini-Projects

A practical suggestion for any reader working through this book systematically: keep every mini-project's code, output, and written reflection in a single, well-organised personal repository, rather than treating each as a disposable classroom exercise. By the time a reader reaches Chapter 18, this repository will contain a genuine, demonstrable body of applied work spanning data cleaning, exploratory analysis, regression and survival modelling, machine learning, spatial analysis, mechanistic disease modelling, and a complete Python outbreak investigation — precisely the kind of portfolio that a prospective employer, PhD supervisor, or research collaborator finds far more convincing than a list of course titles or software proficiencies alone. Structuring this repository with a clear top-level README describing each mini-project in one or two sentences, and linking to the relevant code and output, costs perhaps an hour of additional effort across the whole book but meaningfully increases its long-term professional value.

## A Worked Example

### When Two Reasonable Analysts Disagree, and How to Adjudicate

Two analysts independently examine the same facility-month malaria dataset. Analyst A excludes facility-months with fewer than 10 tests performed, arguing these produce unstable TPR estimates; Analyst B retains them, arguing exclusion introduces selection bias against small, often rural, facilities. Both positions are individually reasonable, and this is a genuinely common situation in applied health data science rather than a sign that one analyst has made an error. The correct way to adjudicate is not to seek a single 'right' answer but to make both approaches explicit and run the key analysis under both, checking whether the substantive conclusion changes. If both approaches converge on the same policy-relevant conclusion, the disagreement is methodologically interesting but practically immaterial; if they diverge, the disagreement itself is the finding, and it should be reported transparently rather than resolved by silently picking whichever approach produces the more convenient result.

## Refining the Inspection Code

### Defensive Programming for Production Use

The inspection code in Section 1.6 is a reasonable first pass, but a version intended for repeated, unattended use (for example, run automatically every month against a new HMIS export) should fail loudly and informatively rather than silently producing a wrong or incomplete result when the input does not match expectations. Two specific improvements are worth making standard practice: validating that expected columns actually exist before referencing them, and returning a structured summary object rather than only printing to the console, so that the result of the check can itself be logged, tested, or fed into a downstream automated pipeline (a pattern extended further in Section 4.13's validation rules).

### *PYTHON example*

```
# Python: a defensive, reusable dataset inspection function
from dataclasses import dataclass, field

@dataclass
class InspectionReport:
    n_rows: int
    n_cols: int
    missing_columns: list = field(default_factory=list)
    implausible_tpr_rows: int = 0
    ok: bool = True

def inspect_malaria_dataset(df, required_cols=("tests_performed", "confirmed_positive")):
    """Validate a facility-month malaria dataset before any analysis proceeds.

    Returns an InspectionReport rather than printing directly, so the
    result can be logged, unit-tested, or used to halt a pipeline.
    """
    missing = [c for c in required_cols if c not in df.columns]
    if missing:
        return InspectionReport(len(df), len(df.columns), missing_columns=missing, ok=False)

    tpr = df["confirmed_positive"] / df["tests_performed"].replace(0, float("nan"))
    implausible = int(((tpr < 0) | (tpr > 1)).sum())
    return InspectionReport(len(df), len(df.columns), implausible_tpr_rows=implausible,
                            ok=(implausible == 0))

report = inspect_malaria_dataset(malaria_df)
if not report.ok:
    raise ValueError(f"Dataset failed inspection: {report}")
```

### *R example*

```
# R: the equivalent defensive inspection function
inspect_malaria_dataset <- function(df, required_cols = c("tests_performed", "confirmed_positive")) {
  missing_cols <- setdiff(required_cols, names(df))
  if (length(missing_cols) > 0) {
    stop("Missing required columns: ", paste(missing_cols, collapse = ", "))
  }
  tpr <- df$confirmed_positive / ifelse(df$tests_performed == 0, NA, df$tests_performed)
  implausible <- sum(tpr < 0 | tpr > 1, na.rm = TRUE)
  list(n_rows = nrow(df), n_cols = ncol(df), implausible_tpr_rows = implausible,
       ok = implausible == 0)
}

report <- inspect_malaria_dataset(malaria_df)
if (!report$ok) stop("Dataset failed inspection: ", report$implausible_tpr_rows, " implausible rows")
```

## Chapter Summary

- Data science for health is a connected pipeline — from problem formulation to decision — not a bag of disconnected techniques.
- Health data carries distinctive constraints: ethical oversight, structural missingness, mixed structured/ unstructured formats, and high-stakes consequences for error.
- Nigerian health data sources used throughout this book include DHIS2/NHMIS, NDHS, malaria diagnostic surveillance, NTD (leishmaniasis) surveillance, adolescent health cohorts, respondent-driven samples, and clinical cohort studies.
- Comparing two seemingly equivalent data streams (e.g., HMIS-reported vs. archive-verified test positivity rate) can reveal systematic bias that only careful attention to data provenance uncovers.
- The first step of any analysis is a basic, scripted inspection of dataset dimensions, types, and plausibility — before any statistical model is fitted.

**Mini-Project**

**Mini-Project 1: Building a Data Provenance Log**

*Choose any dataset you currently work with (or the simulated hmis_malaria_monthly.csv used in this chapter). Produce a one-page 'data provenance log' documenting its full lifecycle, and use it to identify one concrete risk of a naive analysis.*

**Step 1.** State, in one sentence, the exact question the dataset is meant to help answer.

**Step 2.** For each variable, record: how it was captured (paper form, electronic device, laboratory instrument), by whom, and at what frequency.

**Step 3.** Identify at least one point in the data's journey (collection, transcription, aggregation, transmission) where systematic — not random — error could be introduced.

**Step 4.** Using the R or Python code from Section 1.6 as a template, load the dataset and report its dimensions, variable types, and any implausible values you find.

**Step 5.** Write two to three sentences explaining how the risk identified in Step 3 could bias a naive comparison (e.g., a before/after comparison, or a comparison across facilities/states).

**Deliverable:** *A short written provenance log (half a page) plus the annotated inspection code and its output.*

## End-of-Chapter Exercises

1. List three sources of routinely collected health data available in your own state or institution. For each, identify one plausible way its data-generating process could bias a naive analysis.
2. A colleague reports that ‘malaria cases rose 40% this quarter’ using only HMIS aggregate totals. Write three questions you would ask before accepting this claim.
3. Sketch the data science pipeline (Section 1.2) for a question of your choice drawn from your own research area, naming a specific Nigerian data source at the collection stage.
4. Using the worked example in Section 1.5, calculate the percentage-point difference in TPR if the register audit had instead found 430 tests performed and 215 confirmed positive. Comment on whether the direction of the HMIS bias is the same as in the original example.
5. Explain, in two or three sentences, why ‘a large sample size’ does not by itself protect an analysis from the kind of systematic bias illustrated in Section 1.5.
6. Run the R or Python code in Section 1.6 on a dataset of your choice (or a simulated one) and report the number of rows, columns, and any implausible TPR values found.
7. A donor agency asks you to justify why a full data provenance log (Mini-Project 1) is worth the time it takes to produce, given that ‘the numbers are already in the spreadsheet.’ Write a short, persuasive paragraph in response.
8. Describe, in your own words, how the pipeline in Section 1.2 is better thought of as a loop than a straight line, using an example of when a later stage might send you back to an earlier one.

**9.** Which of the three professional roles described in Section 1.3 most closely matches your own background, and which of the other two skill sets do you most need to strengthen?

**10.** Explain briefly why reading a chapter's worked example before attempting its exercises (Section 1.7) might be less effective for long-term learning than attempting the exercises first.

**11.** In the Section 1.8 case study, explain in your own words why Analyst B's estimate, though less dramatic, was ultimately the more useful one for the Ministry of Health.

**12.** Identify which specific step in Analyst B's investigation (reporting completeness, campaign timing, or seasonal decline) you think would be easiest to check first with the least additional data, and explain why.

**13.** If reporting completeness had instead stayed constant across the pre/post period in Section 1.8, would you expect the seasonal-decline confound to fully explain the observed fall on its own? Explain your reasoning.

**14.** Describe one way a health programme could design its own monitoring so that Analyst A's kind of naive before-after comparison is less likely to be misleading in the first place.

**15.** Write a two-sentence explanation, suitable for a non-technical programme manager, of why 'the numbers went down after we started the programme' is not by itself proof the programme caused the improvement.

**16.** Using the four-question checklist in Section 1.9, apply it briefly to a dataset you have worked with recently, writing one short answer to each question.

**17.** Explain why answering these four questions in writing, rather than just 'thinking about it,' makes a meaningful difference in practice.

**18.** Which of the four checklist questions in Section 1.9 do you think is most often skipped in real practice, and why?

**19.** Propose one way a team lead could build this habit into a team's standard workflow, rather than relying on each individual analyst to remember it.

**20.** Explain briefly why this habit is described as something that becomes 'automatic' over time rather than remaining an effortful extra step indefinitely.

**21.** Using Section 1.10, describe a situation where waiting for a perfectly validated analysis would itself cause harm, and explain what a reasonable compromise would look like.

**22.** Explain the difference between a limitation severe enough to invalidate a conclusion outright and one that can be managed with an explicit caveat, using an example.

**23.** Why does this book argue that calibrated judgment, not a fixed rule, is the right way to balance rigour and timeliness?

**24.** Propose one way an analyst could flag 'best current evidence, still being validated' clearly to a decision-maker without either overstating or hiding the analysis's limitations.

**25.** How might this balance differ between a routine annual report and an active outbreak response? Explain your reasoning.

**26.** Summarise, in a single paragraph, the single most important idea from Chapter 1 that you would want a brand-new research assistant to understand on their first day.

**27.** Explain how the malaria TPR example in Section 1.5 and the IRS campaign case study in Section 1.8 both illustrate the same underlying lesson, despite being different scenarios.

**28.** Why does this chapter argue that programming skill alone, without epidemiological or provenance awareness, is not sufficient for responsible health data science?

**29.** Propose one question you would ask a data provider before accepting a new dataset for analysis, based on this chapter's arguments.

**30.** If you had to remove one section from this chapter to make room for something else, which would you keep no matter what, and why?

**31.** In the disagreement between Analyst A and Analyst B (Section 1.12), propose a third possible approach neither analyst considered, and explain its trade-offs.

**32.** Explain why 'running the analysis both ways and checking if the conclusion changes' is a more useful response to analyst disagreement than trying to determine a single objectively correct choice.

**33.** Describe a disagreement you have witnessed or could imagine among colleagues analysing the same dataset, and how Section 1.12's adjudication approach would apply.

**34.** Why might silently picking whichever approach gives the more convenient result be especially tempting, and especially dangerous, under time pressure?

**35.** If your reported finding depended entirely on one debatable analytic choice (such as an exclusion threshold), what should you disclose in your written report?

# 2 The Nigerian Health Data Ecosystem

## A Map of the Terrain

Before a single line of analysis code is written, a Nigerian health data scientist must know the landscape of data sources available, because each has its own unit of observation, reporting frequency, coverage, and known limitations. This chapter maps that terrain, distinguishing routine health information systems from periodic surveys, disease-specific surveillance platforms, and primary research data.

| Data source | Unit of observation | Typical frequency | Primary custodian |
|---|---|---|---|
| DHIS2 / National Health Management Information System (NHMIS) | Health facility (aggregated) | Monthly | Federal/State Ministries of Health |
| Nigeria Demographic and Health Survey (NDHS) | Household / individual | ~Every 5 years | National Population Commission, ICF |
| Nigeria Multiple Indicator Cluster Survey (MICS) | Household / individual | ~Every 4-5 years | National Bureau of Statistics, UNICEF |
| NCDC surveillance (SORMAS, IDSR) | Case / event | Continuous (weekly) | Nigeria Centre for Disease Control |
| Global Fund malaria M&E data (RDT/microscopy/PCR registers) | Health facility / patient test | Monthly | National Malaria Elimination Programme, State programmes |
| Institutional/clinical cohort studies (e.g., UCH, teaching hospitals) | Patient / participant | Study-specific | University research units, hospitals |
| Vital registration (birth/death) | Individual event | Continuous, under-reported | National Population Commission |

## Routine Health Information Systems

### DHIS2 and NHMIS

DHIS2 is the software platform underlying Nigeria's NHMIS, aggregating monthly indicator counts (for example, confirmed malaria cases, antenatal first visits, or DPT3 immunisations) from facility level up through LGA, state, and national levels. Its great strength is coverage and timeliness relative to surveys; its great weakness is that it inherits every error made at the point of paper or electronic entry in the facility, and it reports counts, not denominators drawn from an independent population frame, which means indicators such as coverage rates depend on population projections that can drift substantially from the true catchment population, especially in areas with displacement or rapid urban growth.

A recurring, well-documented pattern in Nigerian malaria data is that the test positivity rate calculated directly from monthly HMIS summary reports tends to run higher than the TPR recomputed from source diagnostic registers and laboratory archives at the same facilities. This gap is not random noise; it typically reflects incomplete transcription of negative test results, selective reporting emphasis on confirmed positive cases, and monthly aggregation choices. A data scientist working with HMIS/DHIS2 exports should therefore always ask: has this indicator been validated against a source register (a Lot Quality Assurance Sampling or Data Quality Assessment exercise), and over what date range?

## Household and Population-Based Surveys

### NDHS and MICS

Where HMIS captures what happens inside the formal health system, the NDHS and MICS are designed to describe the whole population — including those who never visit a facility — through

a stratified, multi-stage cluster sample of households. These surveys are the backbone of Nigeria's estimates for indicators such as under-five mortality, stunting prevalence, contraceptive use, and HIV prevalence, and they are essential for triangulating against HMIS-derived estimates. Their limitations are the mirror image of HMIS: they are infrequent (a five-year gap between rounds means a survey estimate can be badly out of date for a fast-moving indicator), and their cluster design means that a naive analysis that ignores sampling weights and clustering will produce standard errors that are too small and can lead to false statistical significance — a point developed formally in Chapter 7.

## Disease Surveillance Platforms

The Nigeria Centre for Disease Control (NCDC) operates the Integrated Disease Surveillance and Response (IDSR) system and the Surveillance Outbreak Response Management and Analysis System (SORMAS) for case-based reporting of priority diseases such as Lassa fever, cholera, measles, and mpox. Unlike aggregate HMIS counts, these systems capture line-listed, case-level data (in principle allowing analysis of case characteristics, geographic clustering, and time-to-detection), but they are also the systems most sensitive to reporting fatigue during a prolonged outbreak and to inconsistent case definitions across states.

Neglected tropical disease (NTD) surveillance — for example, for cutaneous and visceral leishmaniasis — sits somewhere between the routine and the research categories: baseline mapping exercises produce facility- or LGA-level case counts that can be modelled geographically (Chapter 14) and temporally (Chapter 10), but coverage is typically far patchier than for malaria, because diagnostic capacity for NTDs is concentrated in a small number of referral centres.

## Primary Research and Cohort Data

Much of the most methodologically rich Nigerian health data science happens on purpose-built research datasets: longitudinal adolescent health cohorts tracking mental health and risk behaviours across survey waves, respondent-driven samples (RDS) of key populations such as female sex workers, and clinical studies conducted in partnership with teaching hospitals — for instance, dermatoglyphic (fingerprint pattern) studies comparing children affected by a congenital condition against unaffected controls. These datasets typically have small-to-moderate sample sizes, rich individual-level covariates, and sampling designs (respondent-driven sampling in particular) that require specialised estimators rather than standard survey-weighting approaches; RDS analysis is treated in Chapter 7.

## Triangulation

### Using Multiple Sources Together

The mature practice of Nigerian health data science rarely relies on a single source. A robust analysis of, say, adolescent depression risk will draw on a cohort survey for individual-level modelling (Chapters 8-9) while checking prevalence against MICS; a malaria elimination programme will triangulate HMIS trends against Global Fund diagnostic registers and, where available, community prevalence surveys. This chapter's central lesson, carried forward through the rest of the book, is that the choice of data source is itself a methodological decision with consequences as large as the choice of statistical model.

## Worked Example

### Triangulating Three Sources for a Single Estimate

A State Malaria Elimination Programme wants a single, defensible estimate of malaria test positivity for the 2024 rainy season. Three

candidate figures are available: the HMIS/DHIS2 summary report (68%), the Global Fund diagnostic register audit for a sample of 40 facilities (59%), and a community-based prevalence survey using microscopy in the same LGAs (41%, but measuring prevalence in the general population rather than positivity among those presenting with fever, a materially different quantity — see Section 3.4 on the relationship between prevalence and predictive value). The correct response is not to average these three numbers, which measure different things at different stages of the care-seeking pathway, but to report each with its definition made explicit: HMIS-derived TPR among tested febrile attendees (68%, likely inflated per Section 2.2), register-audited TPR among tested febrile attendees at the sampled facilities (59%, treated as the better estimate of true clinic-level positivity), and community prevalence among the general population regardless of care-seeking (41%, a lower and conceptually distinct figure reflecting the fact that not everyone with parasitaemia seeks care or presents with fever). Presenting all three, correctly labelled, is more useful — and more honest — to a technical working group than collapsing them into one number.

## Practical Data Access

### Working with DHIS2 Exports and Survey Microdata

In practice, DHIS2 data typically arrives to an analyst as a wide-format CSV or Excel export, with one row per facility/period combination and one column per indicator, often requiring reshaping into a long (tidy) format before analysis; NDHS/MICS microdata typically arrives as separate individual-, household-, and birth-recode files in a statistical software format (.dta, .sav) that must be merged on shared household and cluster identifiers. The code below demonstrates the essential reshaping and merging operations a Nigerian health data scientist performs routinely.

*R example*

```r
# R: reshaping a wide DHIS2 export into tidy long format, and merging NDHS files
library(tidyr); library(dplyr); library(haven)

# --- DHIS2: wide to long ---
dhis2_wide <- read.csv("dhis2_export_wide.csv")
dhis2_long <- dhis2_wide %>%
  pivot_longer(cols = starts_with("indicator_"),
               names_to = "indicator", values_to = "value",
               names_prefix = "indicator_")

# --- NDHS: merging individual and household recode files ---
ir_data <- read_dta("NGIR_individual_recode.dta")   # individual recode
hr_data <- read_dta("NGHR_household_recode.dta")    # household recode

ndhs_merged <- ir_data %>%
  left_join(hr_data, by = c("v001" = "hv001", "v002" = "hv002"))  # cluster + household ID
```

## A Note on Data Access Permissions in Practice

Gaining legitimate access to any of the sources catalogued in this chapter typically requires navigating a specific institutional process, and a new Nigerian health data scientist should expect this to take real time: DHIS2/NHMIS access is normally granted through a State or Federal Ministry of Health data request, often requiring an institutional affiliation letter and a stated analytic purpose; NDHS/MICS microdata is available through a formal registration process with the DHS Program or UNICEF MICS website, free for legitimate research use but requiring an account and a brief description of the intended analysis; NCDC surveillance data access is generally restricted to authorised public health staff and formal research collaborations, reflecting its outbreak-sensitive nature; and institutional cohort or clinical data always requires the originating institution's ethical and data-governance approval, per Chapter 16. Building in several weeks of lead time for these approvals, rather than assuming data will be available immediately, is a basic but frequently underestimated project-planning step.

## Extended Case Study

### Reconciling Three Conflicting Population Denominators

A recurring, specifically Nigerian data ecosystem problem deserves its own worked illustration: for almost any LGA, at least three different population figures may be in active use simultaneously — the National Population Commission's projection from the last full census (now more than a decade old and increasingly uncertain in fast-growing or conflict-affected areas), a State Ministry of Health facility catchment-population estimate (often based on an even older assumption and rarely reconciled against migration or displacement), and a survey-based population estimate derived from a recent NDHS or MICS cluster sample covering only a fraction of LGAs directly and requiring modelled small-area estimation for the rest. When a data scientist is asked to compute a rate — malaria cases per 1,000 population, DPT3 coverage as a percentage of the eligible birth cohort — the choice of denominator among these three candidates is not a minor technicality; it can change the resulting rate, and therefore the LGA's apparent ranking relative to its neighbours, substantially.

Consider a specific LGA that has experienced significant internal displacement due to insecurity, with a Ministry of Health catchment estimate of 210,000 (based on pre-displacement figures), an NPC projection of 245,000 (extrapolated from the last census with an assumed uniform growth rate that likely does not reflect the actual displacement-driven population loss), and a recent rapid assessment suggesting an actual current resident population closer to 175,000. Using the Ministry catchment figure as the denominator for a reported 350 confirmed malaria cases yields a rate of 1.67 per 1,000; using the NPC projection yields 1.43 per 1,000; using the rapid assessment figure yields 2.00 per 1,000 — a range wide enough to change whether this LGA appears as a below-average, average, or above-average burden area relative to its neighbours purely as a function of which denominator was

chosen, with no change whatsoever in the actual underlying disease burden. The responsible practice, established in this chapter, is to state explicitly which denominator source was used in any reported rate, to flag LGAs (such as this one) where displacement or rapid population change makes the choice of denominator especially consequential, and, where feasible, to report a sensitivity range across the plausible denominators rather than a single unqualified figure.

## A Comparative Summary Table for Quick Reference

Given how frequently this book returns to the distinct strengths and limitations of Nigeria's major health data sources, it is useful to consolidate the comparative judgments scattered across this chapter into a single quick-reference summary, distilling each source's primary strength and primary caution into a form that can be consulted quickly during project planning, without needing to re-read the chapter in full each time.

| Source | Primary strength | Primary caution |
|---|---|---|
| DHIS2/NHMIS | Near-complete facility coverage, monthly timeliness | Denominators depend on ageing population projections; TPR-type indicators can run high |
| NDHS/MICS | Population-representative, independent of care-seeking | Infrequent (4-5 year gaps); requires design-based (weighted) analysis |
| NCDC/SORMAS-IDSR | Case-level detail, near-real-time for priority diseases | Sensitive to reporting fatigue during prolonged outbreaks |
| Global Fund malaria M&E | Cross-validated against source registers | Coverage limited to supported states/facilities |
| Institutional cohort/clinical studies | Rich individual-level covariates | Small-to-moderate sample sizes; site-specific generalisability limits |

## Interoperability Standards

### Why Systems Struggle to Talk to Each Other

A persistent, structural reason Nigerian health data sources are often difficult to combine (beyond the denominator and definitional issues discussed elsewhere in this chapter) is a lack of shared interoperability standards across different data systems — DHIS2, a hospital's own electronic medical record system, and a research team's custom data collection tool may each use different facility identifiers, different disease coding systems, and different variable naming conventions for conceptually identical information. International standards such as HL7 FHIR (for exchanging clinical health data) and standardised facility master lists are gradually being adopted to address this, but a working Nigerian health data scientist should expect, in the near term, to spend real time manually reconciling facility names, disease codes, and variable definitions across systems that were never designed to communicate with each other directly — precisely the kind of fuzzy-matching and careful cross-referencing work introduced in Chapter 4's discussion of near-duplicate detection, now understood as a symptom of a broader, system-level interoperability gap rather than a series of isolated, one-off data entry errors.

## Worked Example

### Estimating a Population Denominator from Two Conflicting Sources

A recurring practical problem in Nigerian health data science is that no single, universally trusted population denominator exists for computing a coverage or incidence rate: the National Population Commission's projections (based on the last full census, now dated) and DHIS2's own facility catchment population estimates (often derived from historical service statistics rather than an independent count) can disagree by 15% or more for a

fast-growing urban LGA. Rather than arbitrarily choosing one, a defensible approach reports the resulting rate under both denominators explicitly, flags the sensitivity of the finding to the choice of denominator, and, where the distinction matters materially for a decision (as it typically will for a coverage target close to a policy threshold), recommends the more recent, methodologically transparent source while documenting exactly why the choice was made — turning what could be an invisible, undocumented assumption into an explicit, auditable methodological decision.

## Metadata Standards

### DHIS2 Indicator Definitions and Why They Matter

Every indicator within a DHIS2 instance has a formal metadata definition specifying exactly which data elements and category combinations feed into it, and this definition can, in practice, be silently modified over time as a system administrator adjusts an indicator's formula to correct an error or accommodate a new reporting requirement. An analyst comparing an indicator's trend across several years without checking whether its underlying metadata definition remained stable throughout that period risks attributing an artefactual jump (caused by a definition change) to a genuine epidemiological shift. Nigerian DHIS2 instances typically maintain a metadata version history accessible through the platform's API, and a careful analyst working with any multi-year DHIS2 trend should, at minimum, confirm that the indicator's defining formula did not change materially across the period under study.

## A More Robust DHIS2 Import Function

The reshaping code in Section 2.8 assumes a well-formed export with a predictable column-naming convention, an assumption that does not always survive contact with a real DHIS2 instance, where

an indicator name can contain irregular characters or an export can include unexpected extra metadata columns. A production-quality import function should isolate exactly the columns it expects to reshape, warn (rather than silently mis-parse) when an unexpected column pattern is encountered, and preserve the facility identifier as an explicit, typed column throughout, so that a downstream join (Section 2.8's NDHS merge, or any later analysis) cannot silently fail due to a type mismatch between, for example, a numeric and a string-typed facility ID.

***PYTHON example***

```
# Python: a more defensive DHIS2 wide-to-long import function
import pandas as pd
import warnings

def import_dhis2_export(path, indicator_prefix="indicator_"):
    df = pd.read_csv(path, dtype={"facility_id": str})
    indicator_cols = [c for c in df.columns if c.startswith(indicator_prefix)]
    if not indicator_cols:
        warnings.warn(f"No columns found with prefix '{indicator_prefix}' — check export
format")
    long_df = df.melt(id_vars=["facility_id"], value_vars=indicator_cols,
                      var_name="indicator", value_name="value")
    long_df["indicator"] = long_df["indicator"].str.replace(indicator_prefix, "", regex=False)
    return long_df

dhis2_long = import_dhis2_export("dhis2_export_wide.csv")
```

## Chapter Summary

- Nigeria's health data ecosystem includes routine systems (DHIS2/NHMIS), periodic surveys (NDHS, MICS), case-based surveillance (NCDC/SORMAS/IDSR), disease-specific programme data (malaria, NTDs), and primary research/cohort data.
- Each source has a distinct unit of observation, update frequency, and known bias pattern; HMIS-derived indicators such as test positivity rate can differ systematically from source-register-verified values.

- Survey data (NDHS/MICS) requires attention to sampling design (weights, clustering, stratification) before formal inference.
- Triangulating across sources is a core analytic skill, not an optional extra — but triangulation means presenting distinct, clearly-labelled estimates, not averaging quantities that measure different things.
- Practical data access requires reshaping wide DHIS2 exports into tidy format and correctly merging multi-file survey microdata on shared identifiers.

**Mini-Project**

**Mini-Project 2: A One-Page Data Source Comparison Brief**

*Produce a one-page technical brief comparing two Nigerian health data sources for a health indicator of your choice, suitable for a State Ministry of Health technical working group.*

**Step 1.** Select an indicator (e.g., malaria TPR, DPT3 coverage, stunting prevalence) and two data sources that both report on it.

**Step 2.** For each source, state the unit of observation, frequency, and one known limitation, using Section 2.1's table as a template.

**Step 3.** Obtain or simulate a numeric estimate from each source for the same time period and geography.

**Step 4.** Explain, in two to three sentences, any difference between the two estimates in terms of the sources' distinct data-generating processes (not just 'sampling error').

**Step 5.** Conclude with one clear, actionable recommendation for which source (or combination) the technical working group should rely on, and why.

**Deliverable:** *A one-page written brief with at least one comparison table.*

## End-of-Chapter Exercises

1. For an indicator you work with regularly, identify which of the six data sources in Section 2.1 it comes from, and name one limitation specific to that source.
2. Explain, in your own words, why an HMIS-derived test positivity rate might be biased upward relative to a register-verified value.
3. You are asked to estimate malaria prevalence in a state. List the sources you would triangulate and what each would contribute.

**4.** In the worked example of Section 2.7, explain why averaging the HMIS, register-audit, and community-survey figures would produce a meaningless number.

**5.** A DHIS2 export arrives with one column per month (Jan_2024, Feb_2024, ...). Describe, in words, the reshaping operation needed before this data can be plotted as a time series (see Chapter 10).

**6.** Using the R code in Section 2.8 as a guide, write the equivalent Python (pandas) code to reshape a wide DHIS2 export into long format using pandas.melt().

**7.** Explain why NDHS individual and household recode files must be merged on cluster and household identifiers rather than simply concatenated (stacked) into one file.

**8.** You are told a dataset request will take '6-8 weeks for approval.' Using Section 2.9, explain how this should change your project timeline and what you could do productively while waiting.

**9.** Identify one Nigerian health data source not explicitly listed in Section 2.1's table, and describe its unit of observation and update frequency.

**10.** Explain why an NCDC/SORMAS dataset would typically require a more restrictive access process than an NDHS dataset.

**11.** In the Section 2.10 case study, calculate the percentage difference between the highest (2.00 per 1,000) and lowest (1.43 per 1,000) rate estimates relative to the lowest estimate.

**12.** Propose one practical step a State Ministry of Health could take to reduce the kind of denominator uncertainty illustrated in Section 2.10 for a rapidly changing LGA.

**13.** Explain why simply always using the most recent NPC projection, without further investigation, would not fully

solve the denominator problem illustrated in this case study.

**14.** If you had to choose just one of the three denominators in Section 2.10 to report as a headline figure, which would you choose and why, given the LGA's displacement context?

**15.** Describe how you would flag, in a technical report, that a reported rate is sensitive to denominator choice, without simply omitting the rate altogether.

**16.** Using the summary table in Section 2.11, identify which source you would recommend for a question about population-representative HIV prevalence, and justify your choice.

**17.** Which source in Section 2.11's table would you recommend for near-real-time monitoring of a suspected cholera outbreak, and why?

**18.** Explain why 'primary strength' and 'primary caution' are both necessary parts of a useful quick-reference table, rather than strength alone.

**19.** Propose one additional column you would add to the table in Section 2.11 to make it more useful for project planning.

**20.** If you had to choose only one source from Section 2.11's table to fund and strengthen with limited resources, which would you choose for a state focused on malaria elimination, and why?

**21.** Using Section 2.12, explain why two systems using different facility identifiers can make even a simple data merge difficult, beyond any denominator issue.

**22.** What is HL7 FHIR, according to Section 2.12, and what problem is it intended to solve?

**23.** Why does the chapter describe interoperability gaps as a 'structural,' rather than a one-off data entry, problem?

**24.** Propose one practical workaround a research team could use today, while waiting for full interoperability standards adoption, to merge two systems with mismatched facility names.

**25.** Explain the connection between Section 2.12's interoperability problem and Section 4.7's fuzzy-matching technique.

**26.** Explain why this chapter treats 'which data source to use' as a methodological decision on par with choosing a statistical model.

**27.** Summarise, in your own words, the single biggest practical challenge a Nigerian health data scientist faces when combining two or more of this chapter's data sources.

**28.** Propose one improvement to Nigeria's health data ecosystem (realistic or aspirational) that would most reduce the kinds of problems illustrated in this chapter's case studies.

**29.** Why does this chapter emphasise triangulation over simply trusting whichever single source is easiest to access?

**30.** If you could add one new nationally-representative data source to Nigeria's health data ecosystem, what would it measure, and why?

**31.** Explain, using Section 2.14, how an indicator's metadata definition changing partway through a multi-year period could create an artefactual trend jump.

**32.** What specific check would you perform before trusting a five-year DHIS2 trend line for a single indicator?

**33.** Why might a system administrator have a legitimate reason to change an indicator's formula, even though doing so complicates trend analysis?

**34.** Propose a way to flag, in a published trend chart, that an indicator's definition changed partway through the period shown.

**35.** How does the metadata-versioning problem in Section 2.14 relate to the data dictionary concept introduced in Section 4.8?

# 3 Foundation of Probability and Statistical Thinking for Health Data

## Why Probability Comes Before Statistics

Every statistical statement about health data — a confidence interval on a vaccination coverage estimate, a p-value for a treatment comparison, a predicted probability of a positive malaria test — is a probability statement, and it can only be interpreted correctly by someone who understands the probability model that generated it. This chapter builds that foundation from first principles, using health examples throughout, before the book turns to inference (Chapter 7) and modelling (Chapters 8-15).

## Random Variables and Distributions

A random variable maps the outcome of an uncertain process to a number. In health data, discrete random variables are pervasive: the number of confirmed malaria cases reported by a facility in a month, the number of children in a household under five, the result of a rapid diagnostic test (positive/negative). Continuous random variables include a patient's haemoglobin concentration, systolic blood pressure, or age at first antenatal visit.

Four distributions recur constantly in this book and deserve to be internalised early:

- The Bernoulli/Binomial distribution, for a single or repeated yes/no outcome such as a positive diagnostic test

result — the natural model behind logistic regression (Chapter 9) and test positivity rate calculations.

- The Poisson distribution, for counts of events occurring at some average rate over a fixed interval — the natural model for monthly disease case counts reported through HMIS, and the starting point for the Poisson generalised linear models used in Chapter 10's forecasting of NTD trends.
- The Normal (Gaussian) distribution, which, by the Central Limit Theorem, approximates the sampling distribution of many summary statistics (means, regression coefficients) even when the underlying data are not themselves normally distributed — this is why it underlies most classical confidence intervals and hypothesis tests.
- The Exponential and Weibull distributions, which model waiting times or time-to-event data such as time from HIV diagnosis to treatment initiation, or time to relapse — the foundation of the survival analysis methods in Chapter 10.

## Expectation, Variance, and the Idea of a Statistical Model

The expectation (mean) of a random variable is its long-run average value; the variance measures the spread around that average. A statistical model formalises the claim that an observed variable is a realisation of a random variable with a particular distribution, whose parameters (for example, the Poisson rate of malaria cases per facility per month) we wish to estimate from data. Nearly every technique in this book — from a simple two-sample t-test to a deep neural network — can be understood as a way of estimating the parameters of an assumed (explicit or implicit) statistical model and then quantifying how confident we should be in that estimate.

## Conditional Probability, Bayes' Theorem, and Diagnostic Testing

Conditional probability — the probability of an event given that another has occurred — is the single most important idea in this chapter for clinical and public health interpretation, because it underlies the difference between a diagnostic test's sensitivity and specificity on the one hand, and its positive predictive value (PPV) on the other. Bayes' theorem formalises the link:

$$P(\text{Disease} \mid \text{Positive test}) = [P(\text{Positive test} \mid \text{Disease}) \times P(\text{Disease})] / P(\text{Positive test})$$

Consider a rapid diagnostic test (RDT) for malaria with sensitivity 0.95 and specificity 0.90, deployed during the dry season in an area where true malaria prevalence among febrile presentations has fallen to 5%. Applying Bayes' theorem shows that the positive predictive value is only around 33%, meaning two out of every three positive RDT results in that setting are false positives. This is not a flaw in the test; it is a mathematical consequence of testing in a low-prevalence population, and it is precisely why national malaria case management guidelines emphasise pairing RDT results with clinical judgement and, where available, confirmatory microscopy or PCR — the three-way comparison that recurs as a running example throughout this book's diagnostics-related chapters.

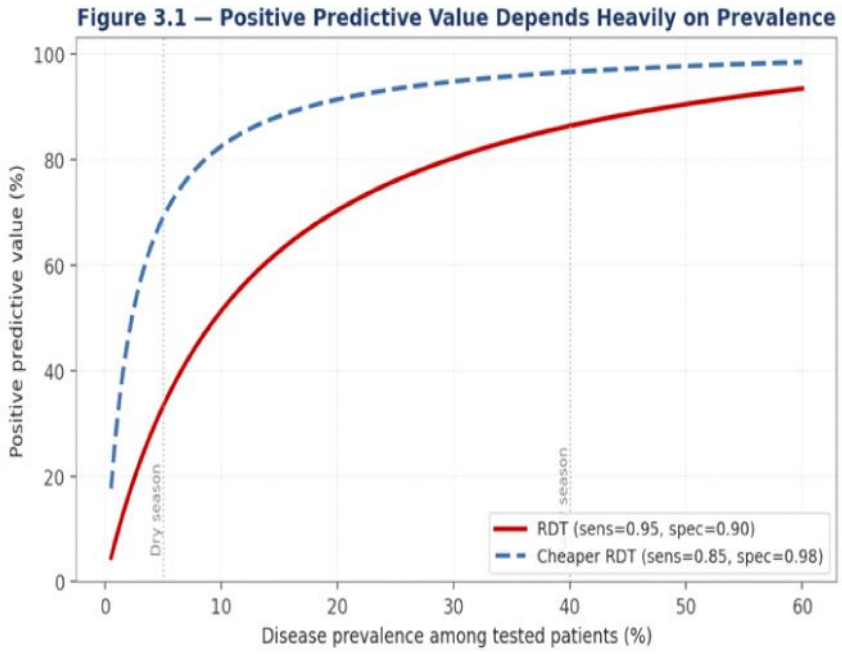



*Figure 3.1 — Positive predictive value as a function of prevalence (Sections 3.4, 3.6, 3.9).*

## Sampling Distributions and the Logic of Uncertainty

A single facility's test positivity rate for a given month is a sample statistic — a snapshot from a random process — and it varies from what one would observe if every febrile patient nationwide could be tested. The sampling distribution describes how that statistic would vary if the process were repeated. Understanding this distinction is what separates a data scientist from someone who merely reads numbers off a dashboard: a facility reporting a TPR of 45% one month and 30% the next has not necessarily seen a real change in transmission; if patient volumes are small, that swing may be entirely consistent with sampling variability, and Chapter 7 develops the formal tools (standard errors, confidence intervals) to make this judgement rigorously rather than by eye.

## Worked Example

### Full Bayes' Theorem Calculation, Step by Step

It is worth working through the malaria RDT calculation from Section 3.4 in full arithmetic detail, because the habit of doing this calculation by hand at least once is what makes the result intuitive rather than merely memorised. Suppose 1,000 febrile patients are tested, and true prevalence among them is 5% (50 truly infected, 950 truly uninfected).

| Group | True status | RDT result | Count |
|---|---|---|---|
| Sensitivity 0.95 applied to 50 infected | Infected | True positive | 47.5 (≈48) |
| | Infected | False negative | 2.5 (≈2) |
| Specificity 0.90 applied to 950 uninfected | Uninfected | True negative | 855 |
| | Uninfected | False positive | 95 |

Total positive RDT results = 48 (true positive) + 95 (false positive) = 143. Positive predictive value = 48 / 143 H" 33.6%, matching the figure quoted in Section 3.4. Repeating this calculation with a rainy-season prevalence of 40% instead of 5% (400 infected, 600 uninfected out of 1,000) gives true positives H" 380, false positives = 60, PPV = 380/440 H" 86.4% — a dramatic illustration of how the same test, with the same sensitivity and specificity, delivers a very different practical reliability depending on the prevalence of disease in the population being tested.

*PYTHON example*

```
# Python: a small reusable function for Bayes' theorem applied to diagnostic testing
def positive_predictive_value(sensitivity, specificity, prevalence):
    true_pos = sensitivity * prevalence
    false_pos = (1 - specificity) * (1 - prevalence)
    return true_pos / (true_pos + false_pos)

for prevalence in [0.05, 0.20, 0.40]:
    ppv = positive_predictive_value(sensitivity=0.95, specificity=0.90, prevalence=prevalence)
    print(f"Prevalence {prevalence:.0%} -> PPV = {ppv:.1%}")
```

## Simulating Distributions to Build Intuition

Simulation is one of the most effective ways to build genuine intuition for a probability distribution, and it will recur throughout this book (notably in Chapter 13's Monte Carlo methods). The code below simulates 10,000 draws from a Poisson distribution with a rate matching a moderate-transmission facility's average monthly case count, and compares the simulated distribution's mean and variance — which should be approximately equal, a defining property of the Poisson distribution exploited in Chapter 8's overdispersion diagnostics.

*R example*

```
# R: simulating a Poisson process and checking mean ≈ variance
set.seed(123)
simulated_cases <- rpois(n = 10000, lambda = 18)  # average 18 cases/month

cat("Simulated mean:", mean(simulated_cases), "\n")
cat("Simulated variance:", var(simulated_cases), "\n")
hist(simulated_cases, breaks = 20, main = "Simulated Monthly Malaria Case Counts",
     xlab = "Cases per month", col = "steelblue")
```

## Independence and the Danger of Assuming It Away

Many of the simplest formulas in introductory probability and statistics — for example, that the variance of a sum of random variables equals the sum of their variances — hold only under an assumption of independence, meaning knowing the value of one variable tells you nothing about another. This assumption is frequently violated in health data in ways that are easy to overlook: repeated measurements on the same patient are not independent of each other (Chapter 9's survival and Chapter 4's longitudinal data structures both hinge on this), and patients treated at the same facility are not independent of each other because they share facility-level influences such as staff training and equipment quality (the same clustering logic that motivates survey design effects in Chapter 7 and grouped cross-validation in Chapter 11). A recurring discipline this book asks of its readers is to explicitly

ask, for any new dataset, what the natural unit of independence actually is, rather than defaulting to treating every row as an independent observation simply because it occupies its own row in a spreadsheet.

## Extended Case Study

### The Base Rate Fallacy Beyond Diagnostic Testing

Section 3.4's diagnostic testing example is the most common illustration of the base rate fallacy in health data science, but the same underlying logical error recurs in a strikingly wide range of other applied settings, and it is worth working through a second, structurally identical example to consolidate the intuition. Suppose a machine-learning-based screening algorithm is developed to flag patient records at high risk of a rare adverse drug reaction, with a reported sensitivity of 90% and specificity of 95%, and suppose the true prevalence of this adverse reaction among all patients prescribed the drug is just 0.5% (a genuinely rare event, typical of many serious but uncommon adverse drug reactions). Applying exactly the same Bayes' theorem calculation as Section 3.6: out of 100,000 prescriptions, 500 will have a true adverse reaction (450 correctly flagged, 50 missed) and 99,500 will not (94,525 correctly cleared, 4,975 incorrectly flagged). The resulting positive predictive value is 450/(450+4,975) H" 8.3% — meaning that of every roughly 12 patient records the algorithm flags as high-risk, only about 1 will actually go on to experience the adverse reaction.

This is not a failure of the algorithm's underlying statistical performance (90% sensitivity and 95% specificity are both genuinely strong figures by ordinary standards) but an unavoidable mathematical consequence of screening for a rare event in a large population — precisely the same base-rate logic as the malaria RDT example, now applied to a pharmacovigilance context rather than a diagnostic test context. The practical implication for a

Nigerian pharmacovigilance or clinical decision-support system is that a low positive predictive value at this stage of algorithm deployment does not necessarily mean the tool should be discarded; it means the tool's output should be explicitly framed and used as a triage/prioritisation signal (worth a closer manual chart review) rather than as a near-certain diagnosis, and that any headline reporting of the algorithm's '90% sensitivity' figure to a non-technical stakeholder audience should always be accompanied by the corresponding positive predictive value at the relevant real-world prevalence, precisely to prevent the base rate fallacy from creating unrealistic expectations about how often a flagged case will turn out to be a true case.

***PYTHON example***

```
# Python: reusing the PPV function from Section 3.6 for a pharmacovigilance screening tool
def positive_predictive_value(sensitivity, specificity, prevalence):
    true_pos = sensitivity * prevalence
    false_pos = (1 - specificity) * (1 - prevalence)
    return true_pos / (true_pos + false_pos)

ppv = positive_predictive_value(sensitivity=0.90, specificity=0.95, prevalence=0.005)
print(f"PPV for rare adverse reaction screening: {ppv:.1%}")   # approx. 8.3%
```

## The Law of Large Numbers and Why it Reassures Us (Sometimes)

The law of large numbers states that as a sample grows, its average value converges toward the true population mean — the formal justification behind the everyday intuition that a larger sample gives a more reliable estimate. This law is frequently invoked, correctly, to argue that a large national survey's headline estimate is more trustworthy than a small facility's monthly figure. It is worth being precise, however, about exactly what the law does and does not guarantee: it guarantees convergence toward whatever the sample is actually drawing from, which is only the true population value if the sampling process itself is unbiased. A very large but systematically biased sample (for example, a very large

dataset drawn only from patients wealthy enough to afford private care) will converge, with great precision, toward the wrong number — precisely the point made about systematic bias in Section 1.5's discussion of sample size versus bias, now given its formal statistical grounding.

## Regression to the Mean

### A Distinct Phenomenon from Bias

A statistical phenomenon easily confused with the systematic biases discussed elsewhere in this chapter, but conceptually distinct, is regression to the mean: an extreme observation (a facility with an unusually high TPR one month, a patient with an unusually high blood pressure reading at one visit) will, on average, be followed by a less extreme observation simply due to the role of chance in producing the original extreme value, even with no genuine underlying change and no systematic bias at all. This matters directly for programme evaluation: if a health programme deliberately targets the facilities or patients with the worst results in a baseline period for a special intervention, some of the improvement observed at follow-up would have occurred anyway, purely from regression to the mean, even if the intervention had no true effect whatsoever — a confound distinct from, and in addition to, the seasonal and reporting-completeness confounds discussed in Section 1.8's IRS campaign case study, and one that specifically arises whenever a group is selected for study or intervention BECAUSE of an extreme baseline value.

## Multiple Testing Revisited

### Why This Isn't Just a Chapter 7 Topic

Chapter 7 formally introduces multiple comparisons corrections, but the underlying probability logic belongs here in the foundations chapter: if a truly null effect has only a 5% chance of appearing 'significant' in any one test, the probability of at least one false

positive across 20 independent such tests is not 5% but 1 " (0.95)^20 H" 64%, a dramatic escalation that follows directly from the basic probability rule for the union of independent events introduced in this chapter's early sections. Internalising this calculation — rather than merely remembering the rule 'correct for multiple comparisons' as an isolated statistical convention — is what allows an analyst to recognise the same underlying risk in contexts far beyond formal hypothesis testing, including, for example, when scanning many exploratory plots (Chapter 5) or reviewing many candidate machine learning models (Chapter 11) for the one that looks best.

## The Gambler's Fallacy and Its Cousins in Health Decision-Making

The gambler's fallacy — the mistaken belief that a run of one outcome makes the opposite outcome 'due' — has direct analogues in health decision-making that are worth naming explicitly. A clinician who has just seen three consecutive malaria-negative RDT results in busy fever season might, without realising it, become subtly more inclined to suspect the fourth febrile patient has malaria 'because we're due for a positive,' when in fact (assuming a stable underlying prevalence) each test result remains statistically independent of the ones before it. This is a specific, avoidable instance of the independence violation discussed in Section 3.8, and recognising it by name — as a specific, well-documented cognitive bias with a name and a known correction — helps a practitioner catch the error in their own reasoning in real time, which is considerably harder to do when the bias remains an abstract, unnamed possibility.

## A More Complete Simulation Function with Reproducibility Guarantees

The simulation in Section 3.7 relies on a global call to 'set.seed()', which works for a single script but becomes fragile once

simulation code is wrapped into a reusable function that might be called many times within a larger analysis, since each call would otherwise draw from wherever the global random state happened to be left by the previous call. A cleaner pattern accepts the seed as an explicit function argument, guaranteeing that any given call is exactly reproducible regardless of what other random operations preceded it elsewhere in the script — a small change that matters considerably once simulation is embedded inside a larger pipeline rather than run as a standalone script.

***R example***

```
# R: a self-contained, independently reproducible simulation function
simulate_case_counts <- function(n = 10000, lambda = 18, seed = 123) {
  old_seed <- if (exists(".Random.seed")) .Random.seed else NULL
  on.exit(if (!is.null(old_seed)) assign(".Random.seed", old_seed, envir = .GlobalEnv))
  set.seed(seed)
  rpois(n, lambda = lambda)
}

# Two independent calls now give identical, reproducible results
identical(simulate_case_counts(seed = 1), simulate_case_counts(seed = 1))  # TRUE
```

***PYTHON example***

```
# Python: the equivalent pattern using a local Generator instance
import numpy as np

def simulate_case_counts(n=10000, lam=18, seed=123):
    rng = np.random.default_rng(seed)   # local, not global — safe to call repeatedly
    return rng.poisson(lam=lam, size=n)

assert (simulate_case_counts(seed=1) == simulate_case_counts(seed=1)).all()
```

## Chapter Summary

- Probability distributions (Bernoulli/Binomial, Poisson, Normal, Exponential/Weibull) provide the generative models underlying nearly every statistical method in this book.
- A statistical model is a claim about which distribution, and which parameters, generated the observed data.

- Bayes' theorem explains why diagnostic test positive predictive value depends heavily on underlying prevalence — a critical concept for malaria and NTD case management.
- Sampling variability means small-facility indicators can swing substantially by chance alone; this must be distinguished from a real change in disease transmission.
- Working through a Bayes' theorem calculation by hand, and simulating a distribution in code, both build the kind of intuition that formula memorisation alone does not provide.

**Mini-Project**

**Mini-Project 3: A Prevalence-Sensitive Diagnostic Calculator**

*Build a small, reusable tool (in R or Python) that reports the positive and negative predictive value of any diagnostic test across a range of prevalence values, and use it to write a one-paragraph clinical recommendation.*

**Step 1.** Extend the Python function in Section 3.6 (or write an equivalent R function) to also compute negative predictive value.

**Step 2.** Run the function across at least five prevalence values spanning a realistic range for a disease of your choice (e.g., 1% to 50%).

**Step 3.** Present the results as a table or simple line plot of PPV and NPV against prevalence.

**Step 4.** Identify the approximate prevalence at which PPV drops below 50% for your chosen test's sensitivity and specificity.

**Step 5.** Write a short paragraph advising a clinician on how to interpret a positive test result differently in a high- versus low-prevalence season.

**Deliverable:** *Working code, one plot or table, and a short written clinical recommendation.*

## End-of-Chapter Exercises

**1.** A malaria RDT has sensitivity 0.95 and specificity 0.90. If prevalence among tested patients is 40% (peak transmission season), compute the positive predictive value and compare it to the 5%-prevalence example in Section 3.4.

**2.** Explain why monthly case counts from a single small health facility are better modelled as Poisson than as Normal.

**3.** A facility's TPR rises from 28% to 41% between two consecutive months with only 20 tests performed each month. Discuss, without doing a formal test, why this alone should not be read as evidence of a transmission increase.

**4.** Using the step-by-step method in Section 3.6, calculate the false negative rate (number of infected people who test negative) when prevalence is 40%, sensitivity is 0.95, and 1,000 patients are tested.

**5.** A new, cheaper RDT has sensitivity 0.85 and specificity 0.98. Using the Python function in Section 3.6, compute its PPV at 5% prevalence and compare it to the original test's PPV. Which property (sensitivity or specificity) matters more at low prevalence, and why?

**6.** Modify the R simulation in Section 3.7 to use a Poisson rate of lambda = 3 (a low-transmission facility). Run it and comment on whether the mean and variance are still approximately equal.

**7.** In your own words, explain the difference between P(Disease | Positive test) and P(Positive test | Disease), and why confusing the two is a common and consequential error in interpreting diagnostic tests.

**8.** Give one example, other than those in Section 3.8, of health data where treating each row as independent would be incorrect, and explain the true unit of independence.

**9.** Explain why the variance of a sum of two correlated random variables is not simply the sum of their individual variances.

**10.** A colleague computes a standard error assuming independence for data that is actually clustered by facility. Will this standard error typically be too large or too small? Explain why.

**11.** Using the pharmacovigilance example in Section 3.9, recompute the positive predictive value if the true prevalence of the adverse reaction were 2% rather than 0.5%, keeping sensitivity and specificity unchanged.

**12.** Explain why a 90% sensitivity, 95% specificity screening tool can still have a very low positive predictive value, without this being a sign the tool is poorly built.

**13.** Propose one way a pharmacovigilance team could communicate the tool's output to clinical staff so that a 'flagged' record is not mistaken for a confirmed adverse reaction.

**14.** In Section 3.9, explain why increasing specificity would generally help the tool's positive predictive value more than a similar increase in sensitivity, given the very low base rate.

**15.** Give one other Nigerian health screening context (besides malaria RDTs and pharmacovigilance) where the base rate fallacy could mislead a non-technical audience, and briefly explain why.

**16.** Explain, using Section 3.10, why a very large sample drawn only from urban, private-hospital patients would not necessarily give a trustworthy estimate of national child mortality.

**17.** What does the law of large numbers guarantee, precisely, and what does it NOT guarantee, according to Section 3.10?

**18.** Give an example of a Nigerian health dataset that could be very large in sample size but still systematically biased in the way Section 3.10 describes.

**19.** Why is 'we have a huge sample size' sometimes used inappropriately to defend a conclusion, in light of Section 3.10?

**20.** Explain briefly how you would investigate whether a large dataset's size is actually protecting it from bias, rather than simply assuming it does.

**21.** Using Section 3.11, explain why some improvement would be expected at follow-up even if an intervention targeting the worst-performing facilities had no true effect at all.

**22.** How is regression to the mean different from the seasonal and reporting-completeness confounds discussed in Section 1.8?

**23.** Propose a study design feature that could help distinguish a genuine intervention effect from regression to the mean.

**24.** Why does selecting facilities specifically BECAUSE of an extreme baseline value make regression to the mean especially likely to occur?

**25.** Explain, in your own words, why regression to the mean is a genuine statistical phenomenon and not evidence of dishonest reporting.

**26.** Explain, in a few sentences, why this chapter argues probability must be understood before statistical inference can be interpreted correctly.

**27.** Summarise the common thread linking the diagnostic testing example (Section 3.4), the pharmacovigilance example (Section 3.9), and the law of large numbers discussion (Section 3.10).

**28.** Why does this chapter emphasise working through calculations by hand at least once, rather than relying solely on software output?

**29.** Propose a scenario from your own field where the base rate fallacy could mislead a well-intentioned analyst.

**30.** What single probability concept from this chapter do you find most useful for your own work, and why?

**31.** Describe, using Section 3.13, a scenario in clinical decision-making where the gambler's fallacy could subtly bias a diagnosis.

**32.** Why is naming a cognitive bias explicitly (rather than leaving it as an abstract possibility) helpful for catching it in real time?

**33.** Explain why three consecutive negative test results do not make a positive result 'due' on the fourth test, assuming a stable underlying prevalence.

**34.** Propose one other context (outside diagnostic testing) where a health professional might unconsciously apply the gambler's fallacy.

**35.** How does the gambler's fallacy relate to, but differ from, the base rate fallacy discussed earlier in this chapter (Section 3.9)?

# Part 2

## Data Wrangling and Exploratory Analysis

# 4 Data Management Types, Sources, Collection, and Cleaning

## Data Types and Structures

Health datasets combine several structural types that must be recognised before any cleaning or analysis begins: cross-sectional data (one observation per subject, as in a single survey round), longitudinal/panel data (repeated observations per subject over time, as in a two-wave adolescent health cohort), and hierarchical/multilevel data (patients nested within facilities nested within LGAs nested within states, as in HMIS data). Getting the structure right determines everything downstream: a longitudinal dataset analysed as if it were cross-sectional will treat correlated repeated measures from the same person as independent, understating uncertainty and, in some cases, reversing the sign of an apparent effect.

Variables themselves are typically classified as nominal (unordered categories, e.g., state of residence), ordinal (ordered categories, e.g., severity grade), interval/ratio (numeric, e.g., haemoglobin level), or count (non-negative integers, e.g., number of ANC visits). This classification determines which summary statistics, visualisations, and models are appropriate — a point revisited throughout Chapters 5, 6, and 9.

## Data Collection Methods

Primary data collection in Nigerian health research spans structured questionnaires (paper-based or electronic, e.g., via ODK/KoboToolbox), clinical measurement and laboratory testing, extraction from patient case notes or registers, and increasingly, respondent-driven and venue-based sampling for hard-to-reach populations. Each method carries characteristic error patterns: interviewer-administered questionnaires are subject to social desirability bias (particularly for sensitive topics such as sexual behaviour or substance use), register extraction is subject to transcription error and missing fields, and electronic data capture, while reducing transcription error, can introduce systematic errors from poorly designed skip logic or default values that respondents never actually confirmed.

## Missing Data

### Mechanisms and Consequences

Not all missing data are equal. Rubin's classic framework distinguishes three mechanisms: missing completely at random (MCAR, where missingness is unrelated to any variable), missing at random (MAR, where missingness depends on observed variables but not the missing value itself), and missing not at random (MNAR, where missingness depends on the unobserved value itself). A patient who defaults from an HIV treatment cohort because their viral load was rising (and therefore they felt too unwell to attend, or became discouraged) is a textbook MNAR scenario — the very outcome of interest drives the missingness — and this is common enough in Nigerian cohort and programme data that analysts should treat MCAR as the exception rather than the default assumption.

- Complete-case analysis (listwise deletion) is only unbiased under MCAR, and even then it discards information and reduces power.

- Single imputation (mean/last-observation-carried-forward) understates uncertainty and can bias estimates, especially LOCF in longitudinal data with a true trend.
- Multiple imputation (e.g., via chained equations, implemented in R's 'mice' package or Python's 'IterativeImputer') is the modern standard under MAR, generating several plausible completed datasets and pooling results to reflect imputation uncertainty.

## Detecting and Handling Data Quality Issues

A systematic data-cleaning workflow checks, in order: (1) structural integrity — do row and column counts match expectations, are identifiers unique; (2) range and plausibility — is a reported age of 200, a temperature of 15°C, or a negative test count present; (3) internal consistency — does a 'pregnant male' or a 'first ANC visit after delivery date' appear; (4) duplicate detection — are the same patient or facility records entered more than once, whether as exact duplicates or as near-duplicates from re-keying; and (5) cross-source reconciliation — for indicators reported in more than one system (e.g., HMIS versus a Global Fund diagnostic register), does the discrepancy fall within an expected tolerance or flag a systemic issue such as the HMIS test-positivity inflation discussed in Chapter 2.

Every cleaning decision of this kind should be logged, not applied silently: a reproducible script (rather than manual editing in a spreadsheet) that documents exactly which rows were dropped, imputed, or flagged, and why, is the single most important habit this chapter can instil, because it is the difference between an analysis that can survive scrutiny — by a supervisor, a journal reviewer, or a Ministry of Health technical committee — and one that cannot.

## Data Management

### Types, Sources, Collection, and Cleaning

***R example***

```
# R: a minimal, auditable cleaning workflow for a facility-level malaria dataset
library(dplyr)

malaria_raw <- read.csv("hmis_malaria_monthly.csv")

malaria_clean <- malaria_raw %>%
  distinct(facility_id, report_month, .keep_all = TRUE) %>%      # drop exact duplicates
  filter(tests_performed >= confirmed_positive,                   # internal consistency
         tests_performed >= 0, confirmed_positive >= 0) %>%
  mutate(tpr = confirmed_positive / tests_performed,
         tpr_flag = tpr > 0.95 | is.nan(tpr))                     # plausibility flag

cat("Rows removed by cleaning:", nrow(malaria_raw) - nrow(malaria_clean), "\n")
cat("Facility-months flagged for review:", sum(malaria_clean$tpr_flag, na.rm = TRUE), "\n")
```

## Data Governance, Storage, and Reproducibility

Health data management does not end with cleaning; it requires a governance discipline: version-controlled raw and cleaned datasets (never overwrite a raw extract), a data dictionary documenting every variable's definition and provenance, de-identification of any dataset before it leaves a secure environment, and, where the dataset involves human subjects, documented ethical approval (from an institutional Health Research Ethics Committee) and a data use agreement where the data originated from a partner institution. These practices are inseparable from the technical skills taught in the rest of this book — a beautifully fitted model built on an unversioned, undocumented dataset is not a reproducible piece of science.

## Worked Example

### Multiple Imputation in Practice

Consider an adolescent health cohort with a depressive symptom score (PHQ-9-type instrument) recorded at baseline and follow-

up, where 15% of follow-up scores are missing, and exploratory analysis (Chapter 5) suggests that missingness is more common among adolescents with high baseline scores and low household food security — a pattern consistent with MAR (missingness explained by observed baseline variables), which multiple imputation can address; if missingness were additionally related to the unobserved follow-up score itself even after conditioning on these variables (e.g., the most severely affected adolescents specifically avoiding follow-up because of their symptoms), the true mechanism would be MNAR, and no amount of standard multiple imputation would fully remove the bias — a sensitivity analysis under plausible MNAR scenarios would then be warranted, a topic returned to briefly in Chapter 9.

***R example***

```
# R: multiple imputation with mice for a longitudinal adolescent cohort
library(mice)

# Predictors: baseline_score, food_security, sex, school_attendance
imp <- mice(cohort_data, m = 20, method = "pmm", seed = 2024,
            printFlag = FALSE)

# Fit the model of interest on each imputed dataset and pool results
fit <- with(imp, lm(followup_score ~ baseline_score + food_security + sex))
pooled <- pool(fit)
summary(pooled)   # pooled coefficients with imputation-adjusted standard errors
```

## Additional Cleaning Patterns in Python

The same cleaning discipline introduced in Section 4.4 applies equally in Python, and analysts moving between languages should be able to translate the underlying logic (not just the syntax) fluently. The example below adds a fuzzy near-duplicate check — useful when facility names have been typed inconsistently across reporting periods (e.g., ‘Eleyele PHC’ versus ‘Eleyele Primary Health Centre’).

***PYTHON example***

```
# Python: cleaning workflow with a fuzzy duplicate check on facility names
import pandas as pd
from rapidfuzz import fuzz

df = pd.read_csv("hmis_malaria_monthly.csv")
df = df.drop_duplicates(subset=["facility_id", "report_month"])
df = df[(df["tests_performed"] >= df["confirmed_positive"]) &
        (df["tests_performed"] >= 0) & (df["confirmed_positive"] >= 0)]
df["tpr"] = df["confirmed_positive"] / df["tests_performed"]

# Flag facility name pairs that are highly similar but not identical (possible re-keying)
names = df["facility_name"].unique()
near_dupes = [(a, b, fuzz.ratio(a, b)) for i, a in enumerate(names)
              for b in names[i+1:] if fuzz.ratio(a, b) > 90 and a != b]
print(f"Potential near-duplicate facility names found: {len(near_dupes)}")
```

## Documenting a Dataset

### The Data Dictionary

A data dictionary — a structured document listing every variable in a dataset along with its definition, units, allowed values or range, and source — is one of the simplest and most under-used tools in applied health data science. For a facility-month malaria dataset, a minimal data dictionary entry for the ‘tests_performed‘ variable would specify: definition (‘total RDT and microscopy tests performed for suspected malaria in the reporting month’), type (integer, non-negative), expected range (0 to roughly the facility’s total monthly outpatient volume), and source (‘HMIS monthly summary form, section 4a’). Producing this document at the start of a project — rather than reconstructing variable definitions from memory months later when a reviewer asks a clarifying question — is a small investment that pays for itself many times over across the life of a project, and it is often a required deliverable for datasets that will eventually be archived or shared with a partner institution.

***R example***

```
# R: auto-generating a skeleton data dictionary from a data frame's structure
library(dplyr)

skeleton_dictionary <- data.frame(
  variable = names(malaria_clean),
  type = sapply(malaria_clean, class),
  n_missing = sapply(malaria_clean, function(x) sum(is.na(x))),
  example_value = sapply(malaria_clean, function(x) as.character(x[1]))
)
write.csv(skeleton_dictionary, "data_dictionary_skeleton.csv", row.names = FALSE)
# The analyst then fills in the 'definition', 'units', and 'source' columns by hand.
```

## Extended Case Study

### Reconstructing a Cleaning Decision Trail After the Fact

A situation this chapter's reproducibility discipline is specifically designed to prevent, but which nonetheless occurs often enough in practice to be worth examining directly: a research team is asked, eighteen months after a paper's publication, to explain why their reported sample size (n=1,842) differs from the raw dataset's row count (n=2,003) held by the partner institution. Because the original cleaning was performed through a combination of manual spreadsheet edits and an undocumented series of ad-hoc filters applied directly in a statistical software's point-and-click interface, no member of the original team can now reconstruct exactly which 161 records were excluded, or why, from memory alone eighteen months later — a serious problem, since a journal or funder inquiry into a discrepancy of this kind can, in the worst case, call the entire paper's findings into question even if the original exclusions were entirely justified.

Contrast this with a hypothetical team that instead followed the scripted, auditable workflow argued for in Section 4.4: their cleaning script, preserved in version control (Chapter 6) alongside a plain-language commit history, shows exactly four exclusion steps applied in sequence — 42 exact-duplicate records removed, 38 records with an implausible test-to-positive ratio flagged and

excluded, 56 records missing the primary outcome variable excluded under a pre-specified complete-case policy, and 25 records from a single facility excluded because that facility's data collection instrument was later found to have used a non-standard case definition for part of the study period. This team can reproduce the exact 1,842 figure in minutes, can show the precise justification for each of the four exclusion categories, and can even quickly re-run the analysis under an alternative exclusion policy (for example, imputing rather than excluding the 56 missing-outcome records, per Chapter 4's multiple imputation guidance) if a reviewer specifically asks for this as a robustness check — a request that would have been effectively impossible to fulfil for the first team without redoing the entire cleaning process from scratch, this time under much greater time pressure and scrutiny.

## A Short Glossary of Cleaning-Stage Terminology

- Listwise deletion: removing an entire record from analysis because at least one of its variables is missing, regardless of whether other variables in that record are complete.
- Winsorising: capping extreme values at a specified percentile (e.g., the 1st and 99th) rather than removing them outright, used to limit the influence of implausible outliers while retaining the record.
- Deduplication key: the combination of variables (e.g., patient ID and visit date) used to identify whether two records refer to the same underlying observation.
- Data lineage: the documented trail showing how a specific value in a cleaned dataset was derived from the original raw source, including every transformation applied along the way.
- Hot-deck imputation: filling a missing value using an observed value from a similar ('donor') record in the same

dataset, an older alternative to the multiple imputation methods emphasised in Section 4.6.

## Metadata

### Documenting the Dataset Beyond Its Variables

Beyond the variable-level data dictionary discussed in Section 4.8, a complete dataset documentation package should include dataset-level metadata: who collected the data and under what ethical approval, the exact date range and geographic scope covered, the sampling or census methodology used, the software and version used for any data collection tool, and a plain-language statement of any known, unresolved data quality concerns at the time of the dataset's release or handover. This metadata is what allows a dataset to be correctly used, years later, by someone who was not part of the original data collection team — without it, even a perfectly clean, well-structured dataset can be silently misapplied to a question or population it was never actually designed to answer, since the misapplication may not be visible from the data's structure alone and requires exactly this kind of contextual documentation to prevent.

## When to Stop Cleaning

### Diminishing Returns and the Risk of Overcorrection

Data cleaning, like any other analytic activity, is subject to diminishing returns, and this chapter closes with a caution rarely stated explicitly in introductory treatments: it is possible to over-clean a dataset, applying an ever-more-aggressive sequence of plausibility filters and imputation adjustments in pursuit of a 'perfectly clean' dataset that no longer resembles the real, imperfect process that generated it. A useful practical discipline is to distinguish clearly, in the cleaning log recommended throughout this chapter, between corrections that fix unambiguous errors (a negative test count, a duplicate record) and adjustments

that reflect a judgement call about plausibility (an unusually high but not impossible TPR); the former should be applied confidently, while the latter should be applied conservatively, documented explicitly, and, where feasible, tested for sensitivity by re-running the key analysis both with and without the judgement-call adjustment.

## Automating Data Quality Checks with Validation Rules

Beyond the one-off cleaning scripts demonstrated earlier in this chapter, a data system that receives new data on an ongoing basis (a monthly HMIS upload, a continuously updated case-based surveillance feed) benefits from automated validation rules that run on every new data submission, immediately flagging an implausible value (a negative count, a TPR outside 0-100%) or a sudden implausible jump relative to the facility's own recent history, before the flawed data ever enters a shared analytic dataset. DHIS2 itself supports configurable validation rules of exactly this kind at the point of data entry, and building the habit of defining validation rules alongside a new indicator — rather than relying entirely on downstream cleaning scripts to catch problems after the fact — shifts data quality assurance from a reactive to a preventive discipline.

***PYTHON example***

```
# Python: a simple automated validation rule applied to each new monthly submission
def validate_submission(row, facility_history):
    errors = []
    if row['confirmed_positive'] > row['tests_performed']:
        errors.append('positive count exceeds tests performed')
    if row['tests_performed'] < 0 or row['confirmed_positive'] < 0:
        errors.append('negative count submitted')
    recent_mean = facility_history['tests_performed'].tail(6).mean()
    if recent_mean > 0 and row['tests_performed'] > 5 * recent_mean:
        errors.append('tests_performed is a 5x jump versus recent 6-month average')
    return errors
```

## A Production-Grade Cleaning Function with Structured Logging

Sections 4.4 and 4.7 print counts of rows affected by each cleaning step, which is adequate for an interactive script but insufficient for a pipeline that runs unattended, where a persistent, timestamped log is needed to audit exactly what happened during each run. Python's built-in 'logging' module (or R's 'logger' package) replaces ad hoc 'print()'/'cat()' statements with structured, leveled log messages that can be routed to a file, filtered by severity, and retained across many pipeline runs — turning the cleaning log recommended throughout Chapter 4 from a one-off console readout into a genuine, persistent audit trail.

***PYTHON example***

```
# Python: a cleaning function with structured logging instead of print statements
import logging

logging.basicConfig(filename="cleaning_log.txt", level=logging.INFO,
                    format="%(asctime)s %(levelname)s %(message)s")

def clean_malaria_dataset(df):
    n_before = len(df)
    df = df.drop_duplicates(subset=["facility_id", "report_month"])
    logging.info(f"Dropped {n_before - len(df)} duplicate facility-month rows")

    n_before = len(df)
    df = df[(df["tests_performed"] >= df["confirmed_positive"]) & (df["tests_performed"] >=
0)]
    logging.info(f"Dropped {n_before - len(df)} internally inconsistent rows")

    df = df.assign(tpr=df["confirmed_positive"] / df["tests_performed"])
    n_flagged = int((df["tpr"] > 0.95).sum())
    logging.warning(f"{n_flagged} facility-months flagged with TPR > 95% for manual review")
    return df
```

## Chapter Summary

- Recognising cross-sectional, longitudinal, and hierarchical data structures determines which analytic methods are valid.

- Missingness mechanisms (MCAR, MAR, MNAR) have different implications; MNAR is common in health cohort/ programme data and complete-case analysis can be badly biased.
- Multiple imputation is the modern standard for handling missing data under MAR; MNAR requires sensitivity analysis rather than standard imputation alone.
- A systematic, scripted (not manual) data-cleaning workflow with an audit trail is essential for reproducibility and defensibility.
- Data governance — versioning, data dictionaries, de-identification, and ethical approval — is part of data science practice, not a separate administrative burden.
- Fuzzy matching helps detect near-duplicate records (e.g., inconsistently typed facility names) that exact-match duplicate checks would miss.

**Mini-Project**

**Mini-Project 4: End-to-End Cleaning Pipeline with an Audit Trail**

*Build a complete, scripted cleaning pipeline for a facility-month dataset (real or simulated) that logs every decision it makes.*

**Step 1.** Simulate or obtain a facility-month dataset with at least 200 rows, deliberately including a few duplicate rows, a few implausible values, and 10-15% missing values in one numeric column.

**Step 2.** Write a script (R or Python) that performs, in order: duplicate removal, plausibility filtering, and missing-value flagging — printing a count of rows affected at each step.

**Step 3.** Apply multiple imputation (Section 4.6) to the flagged missing values and compare the imputed mean to the complete-case mean.

**Step 4.** Write a two-paragraph 'data cleaning report' summarising what was done and why, suitable for inclusion as a methods appendix in a thesis or technical report.

**Step 5.** State explicitly which missingness mechanism (MCAR/MAR/MNAR) you are assuming, and justify it from what you know (or assumed) about the data.

**Deliverable:** *A complete, runnable cleaning script, its printed audit-trail output, and a short written cleaning report.*

## End-of-Chapter Exercises

**1.** Classify each of the following as MCAR, MAR, or MNAR, with justification: (a) a laboratory machine breakdown

causes random missing test results for one week; (b) sicker patients are more likely to miss a follow-up visit; (c) younger respondents are more likely to skip an income question, but income itself doesn't predict skipping once age is accounted for.

**2.** Write out (in pseudocode or R/Python) a cleaning check that would have caught a 'pregnant male' record in a health survey dataset.

**3.** Explain why LOCF (last observation carried forward) can bias a longitudinal analysis of a deteriorating clinical condition.

**4.** In the worked example of Section 4.6, explain in your own words why a sensitivity analysis would be needed if the missingness mechanism were MNAR rather than MAR.

**5.** Using the R code in Section 4.6 as a template, describe (in words) what would change in the 'mice()' call if the outcome variable itself (not just the predictors) had missing values in some imputed datasets but not others.

**6.** Run the Python fuzzy-matching code in Section 4.7 (or describe its logic if you cannot run it) on a list of five deliberately misspelled facility names you invent, and report which pairs are flagged.

**7.** A supervisor asks why you spent two extra days building a scripted cleaning pipeline instead of 'just fixing it in Excel.' Write a two- to three-sentence response drawing on the reproducibility arguments of Section 4.4.

**8.** Using Section 4.8, draft three rows of a data dictionary (variable, definition, type, allowed range) for a dataset you work with regularly.

**9.** Explain why an auto-generated data dictionary skeleton (Section 4.8's R code) still requires manual completion rather than being used as-is.

**10.** A dataset arrives with no documentation at all. Rank, in order of priority, the first three things you would do before beginning any analysis.

**11.** In the Section 4.9 case study, explain what specific piece of information the second (reproducible) team could produce in minutes that the first team could not reconstruct at all.

**12.** Why is 're-running the analysis under an alternative exclusion policy' a valuable capability for responding to a journal reviewer's request, beyond just defending the original figure?

**13.** Describe, in your own words, why manual point-and-click filtering in statistical software creates the same reproducibility risk as manual spreadsheet editing, even though no spreadsheet is involved.

**14.** If you inherited the first (non-reproducible) team's dataset and raw files today, what would be your first practical step toward reconstructing a defensible cleaning trail?

**15.** Explain why 'we are confident the original exclusions were reasonable' is not, by itself, a sufficient response to a reviewer's request to see the exact cleaning steps.

**16.** Using Section 4.10, explain the difference between listwise deletion and winsorising, and give a scenario where each would be more appropriate.

**17.** Why is 'data lineage' (Section 4.10) a stronger standard than simply keeping the raw and cleaned files in the same folder?

**18.** Explain hot-deck imputation in your own words, and describe one advantage multiple imputation (Section 4.6) has over it.

**19.** Propose a deduplication key (Section 4.10) appropriate for a facility-month malaria dataset, and explain why you chose those specific variables.

**20.** Why might winsorising be preferred over deleting outlier records entirely, in a dataset where sample size is already limited?

**21.** Using Section 4.11, list three pieces of dataset-level metadata that a variable-level data dictionary (Section 4.8) alone would not capture.

**22.** Why can a perfectly clean, well-structured dataset still be misapplied without proper metadata, according to Section 4.11?

**23.** Propose a real or plausible scenario where missing dataset-level metadata led (or could lead) to a dataset being used for a population or purpose it was not designed for.

**24.** Why is 'known, unresolved data quality concerns at the time of release' an important item to include in dataset metadata?

**25.** How would you store and share this dataset-level metadata alongside a dataset in your own work?

**26.** Summarise, in your own words, why this chapter treats data cleaning as inseparable from data ethics and governance rather than a purely technical step.

**27.** Explain the common thread linking missing data handling (Section 4.3), duplicate detection (Section 4.4), and metadata documentation (Section 4.11).

**28.** Why does this chapter insist on scripted rather than manual cleaning, even when manual editing might seem faster for a small dataset?

**29.** Propose one cleaning habit from this chapter you intend to adopt in your own work, and explain why you chose it.

**30.** If a colleague asked you to summarise this entire chapter in one sentence, what would you say?

**31.** Explain the difference between the automated validation rules in Section 4.13 and the one-off cleaning scripts introduced earlier in this chapter.

**32.** Why is preventing a data quality problem at the point of entry generally preferable to catching it downstream during cleaning?

**33.** Using the Python code in Section 4.13, describe what additional validation rule you would add for a dataset you work with, and why.

**34.** What is a potential downside of an overly aggressive automated validation rule (for example, one that rejects too many legitimate submissions)?

**35.** How does DHIS2's support for configurable validation rules relate to the interoperability standards discussed elsewhere in this book (Section 2.12)?

# 5 Exploratory Data Analysis and Visualization

## The Purpose of EDA

Exploratory data analysis (EDA), a term coined by John Tukey, is the disciplined practice of summarising and visualising data before committing to a formal model. In health data science, EDA has a specific and non-optional role: it is often the stage at which data quality problems left over from Chapter 4's cleaning process reveal themselves, and it is the stage at which an analyst forms hypotheses about relationships (does severity of adolescent depressive symptoms differ by household food security status? does test positivity vary by rainfall season?) that the formal modelling chapters will later test rigorously.

## Univariate Summaries

For a single variable, the choice of summary depends on its type (Section 4.1). Continuous variables such as haemoglobin concentration are summarised by measures of centre (mean, median), spread (standard deviation, interquartile range), and shape (skewness); the median and IQR are generally preferred over the mean and standard deviation for health measurements known to be skewed, such as clinic wait times or CD4 counts. Categorical variables such as diagnosis category are summarised by frequency tables and proportions, always reported alongside the denominator

(a common error in reporting Nigerian facility-level data is presenting a percentage without stating how many tests or visits it was computed from).

## Bivariate and Multivariate Exploration

Relationships between two variables are explored through cross-tabulations (for two categorical variables), grouped boxplots or violin plots (for a continuous variable across categories, such as birthweight by maternal ANC attendance category), and scatterplots with a smoothed trend line (for two continuous variables, such as rainfall and monthly malaria case counts). Correlation coefficients (Pearson's r for linear association, Spearman's rho for monotonic association robust to outliers) provide a numeric summary but must never be interpreted causally — a point re-emphasised throughout the regression chapters, since a strong correlation between distance to the nearest health facility and undernutrition prevalence does not by itself establish that distance causes undernutrition; both may be driven by underlying poverty.

## Visualization Principles for Health Communication

A visualisation intended for a technical audience (a co-author, a methods reviewer) can carry more statistical detail than one intended for a policy audience (an LGA health officer, a State Ministry of Health briefing). Good practice that applies to both: always label axes and units, always show the denominator behind a rate or proportion, avoid 3-D effects and unnecessary decoration ('chartjunk') that obscure the data, choose a colour palette that remains interpretable to colour-blind readers, and be deliberate about the baseline of a y-axis, since a truncated axis can visually exaggerate a modest change in, for example, immunisation coverage.

***PYTHON example***

```
# Python: visualising monthly malaria TPR trend for a state, with a rainfall overlay
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("state_malaria_monthly.csv", parse_dates=["month"])

fig, ax1 = plt.subplots(figsize=(9, 5))
ax1.plot(df["month"], df["tpr"], color="firebrick", marker="o", label="Test positivity rate")
ax1.set_ylabel("Test Positivity Rate (%)")
ax1.set_xlabel("Month")

ax2 = ax1.twinx()
ax2.bar(df["month"], df["rainfall_mm"], alpha=0.25, color="steelblue", label="Rainfall (mm)")
ax2.set_ylabel("Rainfall (mm)")

fig.suptitle("Malaria Test Positivity vs. Rainfall, [State], 2023-2024")
fig.tight_layout()
plt.savefig("tpr_rainfall_trend.png", dpi=200)
```

## EDA as Hypothesis Generation, Not Hypothesis Confirmation

A discipline this chapter insists on: patterns found through open-ended exploration of a dataset should not be reported with the same confidence as a pre-specified hypothesis test, because searching many possible relationships in a dataset inflates the chance of finding an apparently striking but spurious pattern (a manifestation of the multiple comparisons problem formalised in Chapter 7). Good practice is to treat EDA findings as candidate hypotheses to be confirmed — ideally on an independent dataset or a held-out subsample — rather than as conclusions in their own right.

## Worked Example

### A Complete Univariate-to-Multivariate EDA Walkthrough

Consider a cleaned dataset of 500 adolescents from a two-wave cohort, with variables for depressive symptom score (continuous, 0-27), household food security (categorical: secure/mildly insecure/severely insecure), sex, and school attendance (yes/no).

A complete EDA walkthrough proceeds: (1) univariate — histogram of the symptom score reveals a right-skewed distribution with a long tail of severely affected adolescents, so the median (score 6) and IQR (3-11) are reported rather than the mean; (2) bivariate — a grouped boxplot of symptom score by food security category shows a visibly higher median score in the severely insecure group, and a cross-tabulation of school attendance by food security category shows lower attendance in the severely insecure group; (3) multivariate — a scatterplot of baseline against follow-up symptom score, coloured by food security category, suggests that the food-security gap present at baseline widens by follow-up rather than narrowing. None of these three observations is yet a tested finding — they are candidate hypotheses that Chapter 8's regression modelling (adjusting for sex, school attendance, and baseline score simultaneously) will test formally.

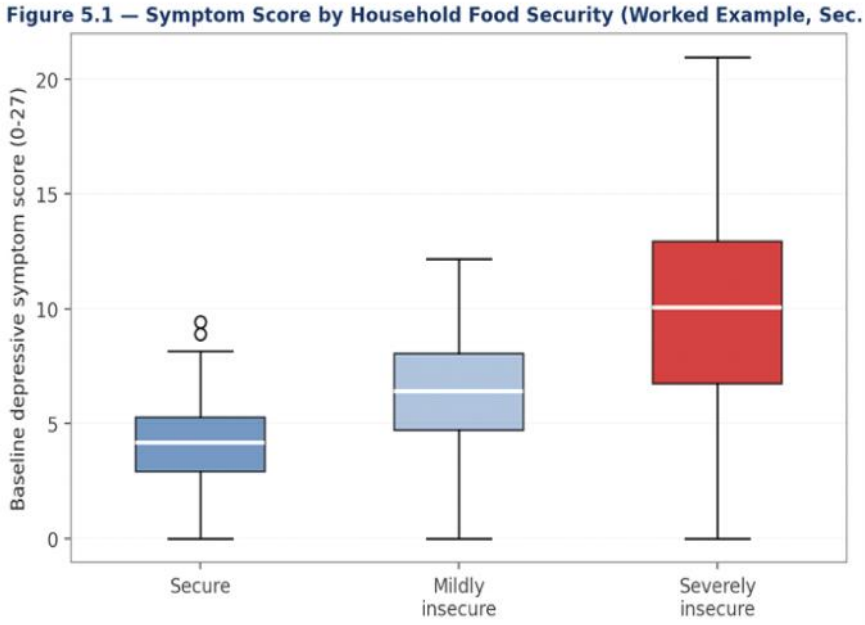



*Figure 5.1 — Depressive symptom score by household food security (Section 5.6 worked example).*

***R example***

```
# R: a compact EDA walkthrough matching the worked example above
library(dplyr); library(ggplot2)

# Univariate
summary(cohort$symptom_score_w1)
ggplot(cohort, aes(x = symptom_score_w1)) + geom_histogram(bins = 15, fill = "steelblue")

# Bivariate: boxplot by food security
ggplot(cohort, aes(x = food_security, y = symptom_score_w1, fill = food_security)) +
  geom_boxplot() + theme_minimal() + labs(y = "Baseline symptom score", x = "")

# Cross-tabulation
table(cohort$food_security, cohort$school_attendance)

# Multivariate: baseline vs follow-up, coloured by food security
ggplot(cohort, aes(x = symptom_score_w1, y = symptom_score_w2, colour = food_security)) +
  geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE) + theme_minimal()
```

## Common EDA Pitfalls in Practice

- Reporting a mean and standard deviation for a visibly skewed variable (e.g., clinic wait time) rather than the more appropriate median and IQR, giving a misleading picture of a ‘typical’ value.
- Interpreting a strong Pearson correlation as evidence of a linear relationship without first plotting the scatterplot, which can reveal a non-linear pattern (e.g., a threshold effect) or an outlier driving the entire correlation.
- Presenting a percentage change (e.g., ’30% increase in cases’) without the underlying counts, which can dramatically overstate or understate practical importance when the base count is very small (a rise from 3 to 4 cases is technically a 33% increase).
- Treating a pattern found by examining dozens of cross-tabulations as a confirmed finding, rather than as one candidate hypothesis among many that must survive formal testing (Chapter 7) and, ideally, replication.

## Small Multiples: Comparing Many Groups at Once

When exploring a variable across many groups simultaneously — for example, monthly TPR trends across all 36 states plus the FCT — a single overlaid plot with 37 lines quickly becomes unreadable. The 'small multiples' technique instead produces a grid of many small, identically-scaled panels, one per group, allowing a reader to scan for both the overall pattern and any state that stands out as an outlier, without the visual clutter of a single overcrowded chart. This technique, alongside the reusable-function approach of Chapter 6, is one of the most practically useful visualisation habits for a Nigerian health data scientist working with state- or LGA-level panel data, since it scales gracefully to dozens of groups in a way that a single combined plot does not.

*R example*

```
# R: small multiples of TPR trend by state using ggplot2 facets
library(ggplot2)

ggplot(malaria_clean, aes(x = report_month, y = tpr)) +
  geom_line(colour = "firebrick") +
  facet_wrap(~ state, ncol = 6) +
  theme_minimal(base_size = 8) +
  labs(title = "Monthly TPR Trend by State", x = NULL, y = "TPR")
```

## Extended Case Study

### When EDA Reveals a Data Collection Problem, Not a Health Pattern

Perhaps the single most valuable real-world function of exploratory data analysis, more valuable in practice than any individual chart type, is its capacity to reveal that an apparent 'finding' is actually an artefact of how the data was collected — and this extended case study works through a realistic example of exactly that discovery process. An analyst exploring a multi-site adolescent health cohort dataset notices, via a simple bar chart of mean depressive symptom score by data collection site, that one site's mean score (4.2) is dramatically lower than all seven other

sites (ranging from 9.8 to 13.1) — an apparent finding that, taken at face value, might suggest that site's adolescent population is genuinely, and remarkably, less symptomatic than every other site in the study.

Following the EDA discipline of this chapter (treating a striking pattern as a hypothesis to be interrogated, not a finding to be reported), the analyst next examines the distribution of individual item-level responses at the low-scoring site specifically, rather than just its summary mean, and discovers that the response distribution is suspiciously concentrated on the lowest possible response option for nearly every item on the symptom checklist — a pattern much more consistent with a systematic data collection or translation problem (for example, an interviewer at that site consistently anchoring respondents toward the lowest response option, or a translation of the instrument into a local language that inadvertently softened the wording of several items) than with a genuine underlying difference in adolescent mental health at that site. A follow-up conversation with the field team confirms that this particular site had, in fact, used a locally back-translated version of the instrument that had not gone through the same formal validation process as the other sites' materials. Reporting the original site-level difference as a substantive finding — without this EDA-driven investigation — would have produced a published claim about geographic variation in adolescent depressive symptoms that was, in reality, entirely an artefact of instrument translation quality, a mistake that a purely confirmatory analysis (jumping straight to a regression model without this exploratory step) would have had no natural opportunity to catch.

## A Short Checklist for Reviewing Someone Else's Exploratory Plots

Reviewing a colleague's or student's exploratory analysis is a routine task for anyone in a supervisory or peer-review role, and

a short, consistent checklist makes this review faster and more thorough than an unstructured read-through. Useful questions to ask of any exploratory plot presented to you: Does the axis start at zero, or is the baseline chosen in a way that could exaggerate an apparent pattern (Section 5.4)? Is the denominator behind any displayed rate or percentage stated somewhere in the figure or its caption? If a mean is shown, does the underlying distribution's shape (visible in an accompanying histogram, or inferable from context) justify using the mean rather than the median? And, most importantly, is this plot being presented as an established finding, or is it appropriately framed as one exploratory observation among several that still requires formal testing or independent confirmation (Section 5.5)?

## Interactive Exploration Versus Static Reporting

The static charts emphasised throughout this chapter (histograms, boxplots, scatterplots) remain the right choice for a fixed report or publication, but an increasingly valuable complementary tool for the EXPLORATION phase itself is an interactive dashboard (built, for example, with R Shiny or Python's Streamlit/Dash) that allows an analyst — or, more importantly, a non-technical domain expert collaborator — to filter, drill down, and re-plot a dataset in real time without needing to write new code for each new question that arises during a collaborative review meeting. This is particularly valuable in the Nigerian context when reviewing a multi-state or multi-facility dataset together with programme staff who have direct, on-the-ground knowledge of specific facilities or states, since they can immediately flag an implausible or interesting pattern in a specific facility of interest during the meeting itself, rather than needing to submit a follow-up request and wait for a new static chart to be produced and circulated afterward.

## Annotated Exploratory Plots

### Turning EDA into a Communication Tool

While Section 5.1 framed EDA primarily as a private, hypothesis-generating activity for the analyst, a well-annotated exploratory plot can also serve as an effective communication tool for a technical audience, provided it retains explicit statistical caveats rather than being polished into a false appearance of confirmed finding. A practical technique is to add a brief textual annotation directly onto an exploratory plot — for example, a note reading 'pattern observed in n=42 facility-months; not yet tested for statistical or practical significance' directly on a suggestive scatterplot — which preserves the plot's exploratory honesty even as it circulates beyond the analyst's own working notes to a supervisor or a co-author who might otherwise mistake a preliminary pattern for a confirmed result.

## Dashboards vs. Reports

### Choosing the Right Format for the Audience

An interactive dashboard (allowing a user to filter by state, LGA, or time period on demand) and a static written report (presenting a fixed, curated set of findings with accompanying narrative interpretation) serve genuinely different purposes, and this chapter's exploratory analysis skills feed into both, but the choice between them should be deliberate rather than default. A dashboard suits an audience that will return to the data repeatedly with different, evolving questions (a State M&E team monitoring an ongoing programme); a written report suits an audience that needs a specific, interpreted, decision-ready answer to a specific question at a specific point in time (a technical working group deciding whether to adjust a treatment protocol). A common and avoidable mistake is building an elaborate interactive dashboard for an audience that actually needed a two-page interpreted memo, or writing a static report for an audience that will need to re-slice the same data dozens of times over the following months.

## A Reusable EDA Summary Function

Rather than re-typing the univariate and bivariate exploration commands from Section 5.6 for every new dataset, wrapping them into a single reusable function that returns a structured summary — one that can itself be inspected, tested, or rendered into a report template — pays off quickly once EDA becomes a routine, repeated step across many datasets and projects, exactly the same efficiency argument made for reusable reporting functions in Chapter 6.

***PYTHON example***

```
# Python: a compact, reusable EDA summary function
def eda_summary(df, outcome_col):
    summary = {
        "n": len(df),
        "n_missing_outcome": int(df[outcome_col].isna().sum()),
        "median": df[outcome_col].median(),
        "iqr": (df[outcome_col].quantile(0.25), df[outcome_col].quantile(0.75)),
        "skew": df[outcome_col].skew(),
    }
    summary["likely_skewed"] = abs(summary["skew"]) > 1
    return summary

result = eda_summary(cohort_df, "symptom_score_w1")
print(result)  # e.g. flags likely_skewed=True, prompting median/IQR over mean/SD (Section
5.2)
```

## Chapter Summary

- EDA summarises and visualises data to detect quality issues and generate hypotheses before formal modelling.
- Choice of summary statistic (mean/SD vs. median/IQR) should match the variable's distribution and type.
- Correlation does not establish causation; bivariate exploration should be interpreted cautiously.
- Visualisations for policy audiences must be honest about axes, denominators, and scale.
- Patterns discovered through open-ended EDA require independent confirmation before being treated as findings.

- A complete EDA walkthrough moves systematically from univariate, to bivariate, to multivariate exploration, generating — not confirming — hypotheses at each stage.

**Mini-Project**

**Mini-Project 5: A Full EDA Report for a Health Dataset**

*Produce a structured, three-part EDA report (univariate, bivariate, multivariate) for a dataset of your choice, following the pattern of Section 5.6.*

**Step 1.** Select a dataset with at least one continuous outcome and at least two categorical or continuous predictors of interest.

**Step 2.** Produce and interpret at least two univariate summaries, choosing appropriate statistics based on distribution shape.

**Step 3.** Produce and interpret at least two bivariate visualisations (one categorical-by-continuous, one continuous-by-continuous).

**Step 4.** Produce one multivariate visualisation (e.g., a scatterplot coloured or faceted by a third variable).

**Step 5.** List three candidate hypotheses your EDA suggests, explicitly labelling them as hypotheses to be tested (not findings), and state which chapter's method you would use to test each.

**Deliverable:** *A short EDA report (plots plus 300-500 words of interpretation) ending with three clearly labelled candidate hypotheses.*

## End-of-Chapter Exercises

1. Given a facility-level dataset with monthly TPR, propose one univariate, one bivariate, and one multivariate exploratory plot, and state what each is intended to reveal.
2. A news report shows a bar chart of vaccination coverage with the y-axis starting at 70% rather than 0%. Explain how this could mislead a reader, using a concrete numeric example.
3. Why should a relationship discovered by scanning 30 variables for correlations with an outcome be treated differently from a single pre-specified hypothesis test?
4. In the worked example of Section 5.6, explain why the median and IQR were chosen over the mean and standard deviation for the depressive symptom score.

**5.** A report states 'malaria cases in LGA X rose 33% this month.' You learn the actual counts were 3 cases last month and 4 this month. Write two sentences explaining why the percentage figure alone is misleading here.

**6.** Using the R code in Section 5.6, describe what change in the boxplot you would expect if severely food-insecure adolescents had, in reality, both a higher median symptom score and much greater variability than the other groups.

**7.** List two of the EDA pitfalls in Section 5.7 that you have personally observed (or could imagine occurring) in a report you have read or written, and explain briefly how you would correct each.

**8.** Explain why a small-multiples plot (Section 5.8) would be more useful than a single overlaid line chart for comparing TPR trends across all 36 states plus the FCT.

**9.** Propose one dataset from your own work where a small-multiples approach would be more informative than a single combined chart, and explain why.

**10.** What is one disadvantage of small multiples compared to a single combined chart, particularly when comparing exact values between two specific groups?

**11.** In the Section 5.9 case study, explain why looking at the distribution of individual item-level responses (not just the mean) at the low-scoring site was the key diagnostic step.

**12.** Propose one additional check, beyond the item-level distribution, that could have helped confirm or rule out the translation-quality explanation earlier.

**13.** Explain why jumping directly to a regression model comparing sites, without first doing the exploratory work

in Section 5.9, would likely have missed this problem entirely.

**14.** If the low-scoring site's item-level distribution had instead looked similar in shape to the other sites' distributions (just shifted lower), would this change your interpretation? Explain why or why not.

**15.** Describe how you would report this discovered instrument-translation issue in a study's limitations section, in two to three sentences.

**16.** Using the checklist in Section 5.10, review a chart you have seen recently (in a report, news article, or presentation) and identify one question from the checklist it fails.

**17.** Why is 'is the denominator stated' one of the most important questions on this checklist, in your view?

**18.** Explain why a reviewer should ask whether a plot is presented as a finding or a hypothesis, even if the plot itself is technically accurate.

**19.** Propose one additional question you would add to the checklist in Section 5.10 for reviewing a health-related exploratory plot.

**20.** How would you handle giving this kind of critical feedback to a junior colleague or student in a constructive, non-discouraging way?

**21.** Using Section 5.11, explain one advantage an interactive dashboard offers during a collaborative review meeting that a set of static charts cannot.

**22.** Why might a non-technical programme staff member using an interactive dashboard catch an implausible pattern that a data scientist working alone might miss?

**23.** Explain why static charts remain the right choice for a final published report, even if interactive tools were used during exploration.

**24.** Propose one risk of relying only on an interactive dashboard for exploration, without ever producing a static, fixed record of what was found.

**25.** Would you recommend an interactive dashboard for the mini-project in Section 5.9's kind of instrument-translation investigation? Explain your reasoning.

**26.** Explain why this chapter treats EDA as a distinct, essential stage rather than something to skip in favour of going straight to formal modelling.

**27.** Summarise the common thread linking the instrument-translation case study (Section 5.9) and the reviewing-someone-else's-plots checklist (Section 5.10).

**28.** Why does this chapter warn against treating a striking EDA pattern as a finding, even when it looks statistically obvious?

**29.** Propose one exploratory check you will now perform routinely, having read this chapter, that you did not previously perform.

**30.** If you had only one exploratory plot to make before trusting a new dataset, which type would you choose, and why?

**31.** Using Section 5.13, describe a situation from your own work where a dashboard would be more appropriate than a static report, and vice versa.

**32.** What is the risk of building an elaborate interactive dashboard for an audience that only needed a two-page memo?

**33.** Explain why an audience that will 're-slice' the same data repeatedly is better served by a dashboard than a fixed report.

**34.** Propose one hybrid approach that could combine some benefits of both a dashboard and a written report for a specific audience.

**35.** How should the choice between a dashboard and a report affect the EDA techniques (Sections 5.2-5.11) an analyst emphasises when preparing it?

# 6 Programming Foundations in R and Python for Health Data Science

## Choosing (or Combining) R and Python

R and Python are both first-class languages for health data science, and this book uses both deliberately rather than favouring one, because Nigerian research groups, journals, and collaborators use both, and a data scientist who can read and write either is more employable and more collaborative. As a rough guide: R's tidyverse and Bioconductor ecosystems, along with its long-standing dominance in classical biostatistics (survival analysis, mixed models, epidemiological packages), make it a natural default for statistical inference and reproducible reporting (R Markdown/ Quarto); Python's scikit-learn, pandas, and deep learning ecosystems (TensorFlow, PyTorch) make it a natural default for machine learning pipelines and production deployment. Chapters 9-15 of this book give code examples in whichever language best fits the technique, and often both, and Chapter 18 is devoted entirely to a deeper, standalone treatment of Python for disease epidemiology.

## Core Data Structures

The data frame (R's 'data.frame'/'tibble', Python's pandas 'DataFrame') is the central object of health data analysis: a rectangular table where rows are observations (patients, facility-

months, survey clusters) and columns are variables. Mastery of a small set of operations — filtering rows, selecting columns, creating derived variables, grouping and summarising, and joining/merging tables on a shared identifier — covers perhaps 80% of the data manipulation that a working analyst performs.

***R example***

```
# R (tidyverse): computing state-level average TPR by quarter from facility data
library(dplyr); library(lubridate)

state_quarterly <- malaria_clean %>%
  mutate(quarter = paste0(year(report_month), "-Q", quarter(report_month))) %>%
  group_by(state, quarter) %>%
  summarise(mean_tpr = mean(tpr, na.rm = TRUE),
            total_tests = sum(tests_performed),
            n_facilities = n_distinct(facility_id),
            .groups = "drop")
```

***PYTHON example***

```
# Python (pandas): equivalent computation
import pandas as pd

df["quarter"] = pd.to_datetime(df["report_month"]).dt.to_period("Q")
state_quarterly = (
    df.groupby(["state", "quarter"])
      .agg(mean_tpr=("tpr", "mean"),
           total_tests=("tests_performed", "sum"),
           n_facilities=("facility_id", "nunique"))
      .reset_index()
)
```

## Writing Reproducible Analysis Scripts

A script is reproducible when a colleague (or the author, six months later) can re-run it from raw data to final output and obtain identical results. This requires: never manually editing a dataset outside of code; setting a random seed wherever randomness is used (train/test splits, bootstrap resampling, Monte Carlo simulation — Chapter 13); recording package/library versions (via ‘renv‘ in R or a ‘requirements.txt‘/virtual environment in Python); and

structuring a project directory consistently (separate folders for raw data, cleaned data, scripts, and outputs). R Markdown and Jupyter/Quarto notebooks that interleave narrative, code, and output are the standard tools for producing a reproducible analytic report — the kind of artefact expected in a thesis chapter, a journal submission, or a technical report to the FMoH.

## Functions, Iteration, and Avoiding Repetition

A recurring inefficiency in applied health analysis is copy-pasting the same block of cleaning or plotting code for each of 36 states or hundreds of facilities. Writing a function once, and applying it iteratively (via a loop, or more idiomatically via ‘purrr::map()‘ in R or a list comprehension/‘.apply()‘ in Python), is not merely a stylistic preference: it means a bug fix or a methodological refinement need only be made in one place, and it dramatically reduces the risk of an inconsistency creeping in between, say, the Lagos and Kano state analyses in a 13-state Global Fund malaria diagnostics report.

## Worked Example

### A Reusable Reporting Function Applied Across States

Suppose a monthly report must be produced for each of Nigeria’s 36 states plus the FCT, each showing the same three summary figures (mean TPR, total tests, number of reporting facilities) and a small trend plot. Writing this once as a function, and applying it iteratively, both saves time and guarantees that every state’s report is generated by exactly the same logic — eliminating the risk that a manual copy-paste error causes, say, Kano’s report to accidentally use Lagos’s date filter.

*R example*

```
# R: a reusable state reporting function applied iteratively with purrr
library(purrr); library(dplyr); library(ggplot2)

make_state_report <- function(state_name, data) {
  state_data <- data %>% filter(state == state_name)
  summary_stats <- state_data %>%
    summarise(mean_tpr = mean(tpr, na.rm = TRUE),
              total_tests = sum(tests_performed),
              n_facilities = n_distinct(facility_id))
  plot <- ggplot(state_data, aes(x = report_month, y = tpr)) +
    geom_line(stat = "summary", fun = mean, colour = "firebrick") +
    labs(title = paste("TPR Trend:", state_name))
  list(state = state_name, stats = summary_stats, plot = plot)
}

all_reports <- map(unique(malaria_clean$state), make_state_report, data = malaria_clean)
names(all_reports) <- unique(malaria_clean$state)
all_reports[["Kano"]]$stats     # access Kano's summary statistics directly
```

*PYTHON example*

```
# Python: the equivalent reusable function applied via a dictionary comprehension
import matplotlib.pyplot as plt

def make_state_report(state_name, data):
    state_data = data[data["state"] == state_name]
    stats = {
        "mean_tpr": state_data["tpr"].mean(),
        "total_tests": state_data["tests_performed"].sum(),
        "n_facilities": state_data["facility_id"].nunique(),
    }
    return stats

all_reports = {s: make_state_report(s, df) for s in df["state"].unique()}
print(all_reports["Kano"])
```

## Version Control Basics for Analysis Projects

Even a solo researcher benefits from basic version control (Git), because it provides a complete, timestamped history of every change to an analysis script, allows safe experimentation on a separate branch without risking a working version, and is the mechanism by which a supervisor, co-author, or reviewer can be given precise, auditable access to the exact code that produced a

given result. A minimal but sufficient workflow commits changes with a short, descriptive message after each meaningful unit of work (a completed cleaning step, a new model specification), rather than in one enormous commit at the end of a project.

*R example*

```
# Terminal (not R code): a minimal Git workflow for an analysis project
# git init                                    # start version control in a project
folder
# git add cleaning_script.R                   # stage a file
# git commit -m "Add facility-month cleaning and TPR plausibility flag"
# git log --oneline                           # view a compact history of all commits
```

## Debugging

### A Systematic Approach

Every analyst, regardless of experience, spends substantial time debugging code that does not run or that runs but produces an unexpected result. A systematic approach saves far more time than ad-hoc trial and error: first, read the actual error message carefully rather than skimming it, since it usually names the specific line and type of problem; second, isolate the smallest piece of code that reproduces the problem, rather than debugging inside a hundred-line script; third, check your assumptions about the data at the point of failure (print the object's dimensions, type, and a few values) rather than assuming the data looks the way you expect; and fourth, when a result runs without error but looks wrong, work backward from the output through each transformation step, checking intermediate results, rather than staring at the final number and guessing at the cause.

## Extended Case Study

### Refactoring an Ad-Hoc Script into a Maintainable Pipeline

Consider a realistic starting point: a 400-line R script, written under deadline pressure to produce a single technical report, that reads three different data files, applies a long sequence of manual

filtering and recoding steps directly in the main script body, produces six plots via copy-pasted and individually hand-edited ggplot2 calls, and saves a final Word document. The script works, in the narrow sense that it produced the report that was needed at the time, but six months later, when the same report needs to be regenerated with an additional quarter of data, the original author (or a colleague standing in for them) faces a script that is difficult to safely modify: it is unclear which of the many sequential filtering steps depend on assumptions specific to the original date range, the six plots' copy-pasted code makes it easy to update five of them correctly while accidentally leaving the sixth using stale logic, and there is no way to verify that the refactored script still produces the same results on the original data before trusting it on the new, extended data.

Refactoring this script following the principles of Chapter 6 involves several concrete steps: extracting the repeated plot-generation logic into a single parameterised function (Section 6.5's pattern) that takes a date range and a plot title as arguments, rather than six independently hand-edited copies; separating the data loading and cleaning steps into their own function with clearly named intermediate outputs, so that a colleague reviewing the code can see exactly what transformation happens at each stage; and, critically, before making any of these changes, running the original script's own results against the refactored version on the identical original dataset to confirm they produce bit-for-bit identical output — a regression test in the software engineering sense, applied here to a data analysis pipeline, which provides concrete evidence that the refactoring has not silently changed the analysis's substantive results. Only after this verification step does the refactored, function-based pipeline get used to generate the updated report with the additional quarter of data, at which point extending the date range and re-running the six plots requires changing a single date-range argument rather than hand-editing

six separate blocks of copy-pasted code — precisely eliminating the fragility that made the original ad-hoc script risky to maintain.

## Choosing Between a Script and a Notebook

A recurring practical question for a Nigerian health data scientist setting up a new project is whether to work primarily in a plain script (a .R or .py file, run from start to finish) or an interactive notebook (R Markdown/Quarto, or Jupyter). Neither is universally superior: a notebook's ability to interleave narrative text, code, and rendered output makes it excellent for exploratory work (Chapter 5) and for producing a shareable analytic report with embedded explanation, but its non-linear, cell-by-cell execution model can silently produce results that depend on the order cells happened to be run in during an editing session, rather than the order they appear in the file — a subtle reproducibility risk unless the notebook is periodically re-run from a completely fresh session ('restart and run all') to confirm it still produces the same results top to bottom. A plain script, by contrast, always executes in a single fixed order and is generally preferred for production pipelines (Chapter 18's field-ready tools) precisely because it removes this specific risk, at the cost of losing the notebook's convenient interleaving of narrative and output during the exploratory phase of a project.

## Testing Your Own Code

### A Brief Introduction

Beyond the informal, ad-hoc checking implicit in the debugging approach of Section 6.7, a more systematic practice — increasingly common in serious applied data science work, though still under-used in much academic health research code — is writing explicit, automated tests for key functions: small pieces of code that check a function produces the expected, known-correct output for a specific, deliberately simple input where the

correct answer can be verified by hand. For example, a custom function computing test positivity rate with a plausibility flag (Section 4.4) could be tested by feeding it a small, hand-constructed dataset with a known TPR of exactly 50% and confirming the function returns exactly that value, and by feeding it a deliberately implausible case (more positives than tests performed) and confirming the plausibility flag correctly triggers. Writing even a handful of such tests for a project's most important, most-reused functions provides a fast, automatic way to confirm that a later change to the function (a bug fix, a refactoring per Section 6.8) has not silently broken its correct behaviour for the cases that matter most.

## Continuous Integration for Analysis Pipelines

### A Light-Touch Introduction

Continuous integration (CI) — automatically re-running a project's code (and, ideally, a small set of automated checks) every time a change is committed to version control — is standard practice in professional software engineering and is increasingly accessible to health data science projects through free platforms such as GitHub Actions. A light-touch CI setup for a typical analysis project might simply re-run the core cleaning and summary-statistics script on every commit and fail loudly if the script errors or if a key summary statistic (such as total row count) falls outside an expected range, catching an accidental regression (for example, a cleaning script silently breaking after an update to a package it depends on) far earlier than a manual, occasional re-run would.

### Environment Management

### Why 'It Works on My Machine' Is a Warning Sign

A script that runs correctly on the original analyst's computer but fails, or produces subtly different results, on a colleague's machine

is a common and specifically diagnosable reproducibility failure: differing package versions between the two environments. R's 'renv' and Python's virtual environments (or 'conda' environments) address this by recording the exact versions of every package used in a project and allowing any future user (including the original analyst, on a new machine) to recreate an identical computational environment with a single command, rather than manually reinstalling packages and hoping the versions happen to match. Treating 'it works on my machine' as a warning sign rather than a satisfactory endpoint — and routinely capturing the project's environment specification alongside its code — is a small additional step that closes one of the most common and otherwise invisible gaps in analytic reproducibility.

## Adding Unit Tests to the Reporting Function

Section 6.10 introduced testing as a brief orientation; applying it concretely to the reusable reporting function from Section 6.5 shows what a genuinely useful test looks like: not a test that checks the function runs without crashing, but one that checks the function returns the correct numeric answer on a small, hand-constructed input whose correct output the analyst already knows by hand-calculation, catching a logic error (such as an accidental sum instead of a mean) that a mere "does it run" check would miss entirely.

***PYTHON example***

```
# Python: a genuine unit test for the state reporting function (Section 6.5)
import pandas as pd

def test_make_state_report_computes_correct_mean():
    toy_data = pd.DataFrame({
        "state": ["Oyo", "Oyo", "Oyo"],
        "tpr": [0.2, 0.4, 0.6],
        "tests_performed": [100, 100, 100],
        "facility_id": ["F1", "F2", "F3"],
    })
    result = make_state_report("Oyo", toy_data)
    assert abs(result["mean_tpr"] - 0.4) < 1e-9, "Mean TPR calculation is incorrect"
    assert result["n_facilities"] == 3

test_make_state_report_computes_correct_mean()  # raises AssertionError if the logic regresses
```

## Chapter Summary

- R and Python are complementary tools; this book uses both, matched to the strength of each ecosystem, with a dedicated Python chapter (18) at the close of the book.
- The data frame and a small set of manipulation verbs (filter, select, mutate, group/summarise, join) underlie most applied analysis.
- Reproducibility requires scripted (not manual) workflows, fixed random seeds, recorded dependencies, and consistent project structure.
- Functions and iteration prevent the errors and inconsistency that arise from copy-pasted, state-by-state or facility-by-facility code.
- Basic version control (Git) provides an auditable history of an analysis project and supports safe collaboration.

**Mini-Project**

**Mini-Project 6: A Multi-State Automated Reporting Pipeline**

*Build a small pipeline (R or Python) that automatically generates a one-page summary for each of at least five states/groups in a dataset, using a single reusable function.*

**Step 1.** Simulate or obtain a facility-month dataset covering at least five states.

**Step 2.** Write a single function (following Section 6.5) that computes three summary statistics and produces one plot for a given state.

**Step 3.** Apply the function iteratively (via map()/apply(), not copy-paste) across all states, storing the results in a list or dictionary.

**Step 4.** Extract and display the results for two specific states to demonstrate the pipeline works correctly.

**Step 5.** Initialise a Git repository for your project (if not already done) and make at least three meaningful, separately-described commits as you build the pipeline.

**Deliverable:** *The reusable function, its iterative application across states, and a short Git commit history (a screenshot or `git log` output is sufficient).*

## End-of-Chapter Exercises

**1.** Convert a manual, spreadsheet-based process you currently use for a repeated task (e.g., producing a monthly report

for each LGA) into a plan for a scripted, reproducible pipeline.

2. Write an R or Python function that takes a facility-month data frame and returns TPR with a flag for implausible values (see Section 4.4), then apply it across multiple states using iteration rather than copy-paste.
3. Explain, to a non-programmer colleague, why setting a random seed matters for a machine learning pipeline that will be run again next month on updated data.
4. In the worked example of Section 6.5, explain what would go wrong (in terms of reproducibility and correctness) if each state's report were instead produced by copy-pasting and manually editing the filter condition 36 times.
5. Using the Python code in Section 6.5, modify 'make_state_report()' to also return the number of months with zero reported tests, and explain why this additional check might be useful for a state-by-state HMIS data quality audit.
6. Explain, in your own words, the difference between committing all your work in one single Git commit at the end of a project versus committing incrementally after each meaningful step.
7. A colleague says version control is 'only for software engineers, not for a statistics thesis.' Write a short response, drawing on the reproducibility arguments made across Chapters 4 and 6.
8. Using the systematic debugging approach in Section 6.7, describe the first two steps you would take if a script produced an error you did not immediately understand.
9. Explain why isolating the smallest piece of code that reproduces a bug is more efficient than debugging within a full, long script.

**10.** A script runs without any error but produces an obviously wrong summary statistic. Using Section 6.7's fourth step, describe your debugging approach.

**11.** In the Section 6.8 case study, explain why running the refactored pipeline against the original script's results on the SAME original data was an essential verification step, not an optional extra.

**12.** Describe one risk that would remain even after this verification step, when the refactored pipeline is later run on the new, extended quarter of data.

**13.** Why does extracting the six copy-pasted plots into a single parameterised function reduce the risk of the specific 'five out of six updated correctly' failure mode described in Section 6.8?

**14.** If you inherited a 400-line ad-hoc script like the one in Section 6.8, would you refactor the whole script at once or incrementally? Justify your choice.

**15.** Explain, in your own words, what a 'regression test' means in the data-analysis context described in Section 6.8, as distinct from its meaning in Chapter 8's regression modelling.

**16.** Using Section 6.9, explain the specific reproducibility risk introduced by a notebook's non-linear, cell-by-cell execution model.

**17.** Why does 'restart and run all' meaningfully reduce the risk described in Section 6.9, compared to simply saving the notebook after editing?

**18.** For which of the tasks in this book (e.g., exploratory analysis, a field-ready SEIR tool) would you personally choose a notebook, and for which a plain script? Justify your choices.

**19.** Propose one hybrid workflow that gets some of the benefits of both notebooks and scripts for a single project.

**20.** Explain to a student who has only ever used notebooks why a production tool (like the Section 18.11 dashboard) is more often built as a script or application than as a notebook.

**21.** Using Section 6.10, write (in words) two test cases you would use to check a function that calculates case fatality ratio (Section 18.5).

**22.** Why is testing a function with a deliberately implausible input (e.g., more deaths than total cases) useful, in addition to testing with a normal, expected input?

**23.** Explain how automated tests specifically support the refactoring process described in Section 6.8.

**24.** Why are automated tests described as 'under-used' in academic health research code, and what barrier do you think most contributes to this?

**25.** Propose one function from any chapter in this book that you would prioritise writing a test for first, and explain your reasoning.

**26.** Summarise, in your own words, why this chapter treats programming practice (functions, version control, testing) as inseparable from good statistical analysis.

**27.** Explain the common thread linking the refactoring case study (Section 6.8) and the notebook-versus-script discussion (Section 6.9).

**28.** Why does this chapter argue that reproducibility habits benefit even a solo researcher, not just large teams?

**29.** Propose one habit from this chapter (functions, testing, version control, or documentation) you consider most valuable for your own current work, and explain why.

**30.** If you had to teach this chapter's single most important lesson to a new graduate student in one sentence, what would it be?

**31.** Explain, using Section 6.12, why 'it works on my machine' is a warning sign rather than a satisfactory outcome.

**32.** What specific information does an environment specification (via renv or a Python virtual environment) capture that plain code does not?

**33.** Describe a scenario where a package version mismatch between two collaborators' machines could silently produce different analysis results.

**34.** Why might capturing an environment specification be considered as important as version-controlling the code itself?

**35.** If you have never used an environment management tool before, what is the first concrete step you would take to start?

# Part 3

## Statistical Inference and Modelling

# 7 Statistical Inference and Hypothesis Testing in Health Research

## From Sample to Population

Statistical inference is the process of drawing defensible conclusions about a population from a sample, together with a rigorous statement of how confident that conclusion should be. Every quantity computed in Chapter 5's EDA — a mean, a proportion, a difference between groups — is an estimate, and this chapter develops the tools (standard errors, confidence intervals, hypothesis tests) needed to attach a defensible measure of uncertainty to that estimate.

## Confidence Intervals and Hypothesis Tests

A confidence interval expresses a range of plausible values for a population parameter, given the observed sample; a 95% confidence interval, correctly interpreted, means that if the sampling and estimation procedure were repeated many times, approximately 95% of the resulting intervals would contain the true parameter value — it does not mean there is a 95% probability that this particular interval contains the true value. A hypothesis test formalises a comparison against a null hypothesis (typically 'no difference' or 'no association') and produces a p-value: the probability, if the null hypothesis were true, of observing a result

at least as extreme as the one obtained. A small p-value is evidence against the null hypothesis; it is not, by itself, a measure of clinical or public health importance, and this distinction between statistical significance and practical significance is one this chapter insists on throughout — a difference in mean birthweight of 15 grams between two groups can be statistically significant in a very large sample while being clinically negligible.

- Two-sample t-test / Mann-Whitney U test: comparing a continuous outcome (e.g., mean CD4 count) between two independent groups, using the non-parametric Mann-Whitney U when normality is doubtful or sample sizes are small.
- Paired t-test / Wilcoxon signed-rank test: comparing before/after measurements on the same subjects (e.g., pre- and post-treatment schistosomiasis egg counts).
- Chi-squared test / Fisher's exact test: testing association between two categorical variables (e.g., diagnostic test result by season), with Fisher's exact test preferred when expected cell counts are small.
- One-way ANOVA / Kruskal-Wallis test: comparing a continuous outcome across three or more groups (e.g., mean TPR across three ecological zones).

## The Multiple Comparisons Problem

Testing many hypotheses in a single dataset — as happens routinely when comparing an outcome across dozens of LGAs, or screening many candidate risk factors for an outcome — inflates the overall (family-wise) chance of a false positive well beyond the nominal 5% used for any single test. Corrections such as the Bonferroni adjustment (conservative, dividing the significance threshold by the number of tests) or the Benjamini-Hochberg procedure (controlling the false discovery rate, generally preferred when many tests are run, as in genomic or wide screening studies) should

be applied whenever a study design involves testing more than a small, pre-specified number of hypotheses.

## Survey-Weighted and Design-Based Inference

As introduced in Chapter 2, the NDHS and MICS use a complex, stratified, multi-stage cluster design. Ignoring this design — treating the sample as if it were a simple random sample — typically understates standard errors, because individuals within the same cluster (village or enumeration area) tend to be more similar to each other than to individuals in a different cluster (a phenomenon quantified by the intraclass correlation coefficient and summarised in the design effect). Correct analysis uses survey-specific software (R's 'survey' package or Python's 'samplics') that incorporates sampling weights, strata, and primary sampling units, producing standard errors that are typically wider — and therefore more honest — than a naive calculation.

## Respondent-Driven Sampling: A Special Case

Respondent-driven sampling (RDS), used to reach populations without a sampling frame — such as female sex workers or other key populations relevant to HIV programming in Nigeria — recruits participants through peer referral chains rather than random selection, and standard survey-weighting methods do not directly apply. RDS-specific estimators (such as the Volz-Heckathorn estimator, successor-of-successor estimator, and RDS-II variants) adjust for each participant's network size and for homophily (the tendency of peers to recruit similar peers) to produce a population-representative estimate. Comparing multiple RDS estimators on the same dataset — a standard robustness check — often reveals that estimates are sensitive to the assumed recruitment model, which is why a careful analyst reports a range of estimators rather than a single point estimate for RDS-derived prevalence figures.

## Worked Example

### A Full Hypothesis Test, Start to Finish

Suppose we wish to test whether mean birthweight differs between infants of mothers who attended four or more ANC visits (n = 180, mean = 3120g, SD = 410g) and those who attended fewer than four (n = 95, mean = 2980g, SD = 450g). The null hypothesis is that the true population mean birthweight is equal in both groups. An independent two-sample t-test (Welch's version, which does not assume equal variances) gives a test statistic of approximately t = 2.51 with roughly 180 degrees of freedom, corresponding to a two-sided p-value of about 0.013. Because 0.013 < 0.05, we reject the null hypothesis and conclude there is statistically significant evidence of a difference in mean birthweight between the two groups. The 95% confidence interval for the difference (140g) is approximately (30g, 250g). Note, following Section 7.2, that a 140g mean difference — while statistically significant and plausible given known physiology — should also be interpreted for its practical importance: is a 140g average difference large enough to change clinical or programmatic practice? That is a judgement the statistics alone cannot make.

***R example***

```
# R: full two-sample t-test workflow with confidence interval
anc_high <- rnorm(180, mean = 3120, sd = 410)   # simulated for illustration
anc_low  <- rnorm(95,  mean = 2980, sd = 450)

test_result <- t.test(anc_high, anc_low, var.equal = FALSE)
print(test_result)

cat("Mean difference:", round(diff(rev(test_result$estimate)), 1), "g\n")
cat("95% CI:", round(test_result$conf.int[1],1), "to", round(test_result$conf.int[2],1),
"g\n")
```

## Multiple Comparisons in Practice: A 20-LGA Example

Suppose an analyst tests whether DPT3 coverage differs from the national average in each of 20 LGAs, running 20 separate one-

sample tests at alpha = 0.05. Even if there were truly no difference anywhere, we would expect roughly 20 x 0.05 = 1 false positive by chance alone. Applying a Bonferroni correction (Section 7.3) means each individual test must now achieve $p < 0.05/20 = 0.0025$ to be declared significant — a substantially stricter bar — while the Benjamini-Hochberg procedure instead ranks all 20 p-values and compares each to a sliding threshold, generally retaining more true positives than Bonferroni while still controlling the overall false discovery rate. The code below demonstrates both corrections on a simulated set of 20 p-values.

***PYTHON example***

```
# Python: applying Bonferroni and Benjamini-Hochberg corrections to 20 p-values
import numpy as np
from statsmodels.stats.multitest import multipletests

np.random.seed(1)
raw_pvalues = np.random.uniform(0, 1, 20)          # simulated p-values from 20 LGA tests
raw_pvalues[3] = 0.001                              # one genuinely strong signal

bonferroni_reject, bonferroni_p, _, _ = multipletests(raw_pvalues, alpha=0.05,
method="bonferroni")
bh_reject, bh_p, _, _ = multipletests(raw_pvalues, alpha=0.05, method="fdr_bh")

print("Bonferroni significant LGAs:", np.where(bonferroni_reject)[0])
print("Benjamini-Hochberg significant LGAs:", np.where(bh_reject)[0])
```

## Worked Example

### Survey-Weighted vs. Naive Standard Errors

Using a simulated NDHS-like dataset with 30 clusters of 20 households each, suppose the naive (unweighted, unclustered) standard error for estimated stunting prevalence is 1.2 percentage points, but the correct, design-based standard error (accounting for clustering and sample weights) is 2.1 percentage points — a design effect of roughly (2.1/1.2)^2 H” 3.1, meaning the effective sample size is about one-third of the nominal sample size. A 95% confidence interval built on the naive standard error would be misleadingly narrow, potentially declaring a year-on-year change

in stunting prevalence 'significant' when the correctly-computed interval shows it is well within the range of sampling variability.

***R example***

```
# R: survey-weighted estimation using the survey package
library(survey)

ndhs_design <- svydesign(ids = ~cluster_id, strata = ~state,
                         weights = ~sample_weight, data = ndhs_data)

svymean(~stunted, ndhs_design)        # design-based mean and standard error
# Compare to the naive (incorrect) calculation:
mean(ndhs_data$stunted); sd(ndhs_data$stunted) / sqrt(nrow(ndhs_data))
```

## Effect Sizes

### Going Beyond Significance

Alongside a p-value, an effect size — a standardised measure of the magnitude of a difference or association, independent of sample size — allows results to be compared across studies and gives a more complete picture than significance alone. Cohen's d (for a difference in means, expressed in standard deviation units) and the odds ratio or relative risk (for a difference in proportions) are the effect sizes most commonly reported in Nigerian health research; conventional (though somewhat arbitrary) benchmarks describe a Cohen's d of 0.2 as small, 0.5 as medium, and 0.8 as large, but the appropriate benchmark for 'clinically meaningful' should ultimately be set by domain knowledge of the outcome in question, not by these generic statistical conventions alone.

## Extended Case Study

### A Non-Significant Result That Was the Right Answer

This chapter has repeatedly emphasised that statistical significance is not the same as practical importance; it is equally important to internalise the reverse lesson, that a properly conducted, non-significant result can itself be the correct and most useful answer to a research question, rather than a disappointing failure to be

explained away. Consider a study testing whether a low-cost, community health worker-delivered adherence counselling intervention improves viral suppression rates among a cohort of 240 adults on antiretroviral therapy, compared to a control group of 235 receiving standard care. The observed difference in 12-month viral suppression is small (74% intervention vs. 71% control) and the associated 95% confidence interval for the difference (-4 to +10 percentage points) comfortably includes zero, giving a non-significant result ($p = 0.29$) by any conventional threshold.

A team under pressure to report a 'positive' finding might be tempted to search for a subgroup in which the effect appears larger (a form of the multiple-comparisons problem discussed in Section 7.3, sometimes called 'p-hacking' when done without pre-specification and full disclosure), or to describe the result using softened language such as 'a promising trend toward improved suppression' that overstates the actual evidence. The methodologically correct response, and the one this book argues for throughout, is instead to report the result plainly: this study did not find statistically significant evidence that the counselling intervention improved viral suppression, the confidence interval is consistent with anything from a modest harm to a moderately large benefit, and a substantially larger sample size would be needed to distinguish a true small-to-moderate effect from no effect at all. This honestly-reported null result remains genuinely valuable to the HIV programme deciding whether to scale up the intervention nationally: it correctly tells them that the evidence so far does not justify the additional cost of scale-up, while leaving open (via the confidence interval, not via speculative subgroup mining) the possibility that a moderate true benefit exists and a larger, adequately powered follow-up study would be needed to detect it with confidence.

## One-Sided Versus Two-Sided Tests

This chapter's hypothesis tests have all been presented as two-sided, testing for a difference in either direction, which is the appropriate and more conservative default for most health research questions. A one-sided test, which tests only for a difference in one pre-specified direction (for example, testing only whether a new bednet reduces, rather than testing for any difference in either direction from, malaria incidence), produces a smaller p-value for the same data and is therefore easier to declare significant — a property that makes one-sided tests appropriate only when a difference in the opposite direction would be scientifically implausible or practically irrelevant, and when this choice is made and stated BEFORE seeing the data, never chosen after the fact simply because it produces a more favourable p-value for a result already observed.

## Sample Size and Power Calculations Before Data Collection

This chapter's methods for interpreting a completed study's results are only half of the statistical inference picture; before data collection begins, a power calculation determines the sample size needed to have a reasonable chance (conventionally 80% or 90%) of detecting a specified minimum effect size, if that effect genuinely exists, at a chosen significance level. Skipping this step is a common and consequential planning error: a study that turns out to be underpowered for the effect size it was actually looking for will produce the kind of wide, inconclusive confidence interval discussed in Section 7.10's HIV counselling example — not because the intervention necessarily has no effect, but because the study was never designed with enough participants to distinguish a real, moderate effect from no effect at all. Performing a power calculation at the design stage, using a realistic, ideally literature-informed estimate of the minimum clinically important effect

size, is therefore a critical planning step that determines whether a study's eventual result — significant or not — will actually be informative.

## Equivalence Testing

### When 'No Significant Difference' Isn't Enough

A standard hypothesis test, as developed in this chapter, is designed to detect a difference from a null value; failing to reject the null hypothesis is not the same as positive evidence that two groups are truly equivalent, since a study may simply lack the statistical power to detect a real but modest difference (Section 7.12). When a research question genuinely requires evidence of equivalence — for example, showing that a cheaper, locally-produced diagnostic reagent performs no worse than an imported reference reagent, within a pre-specified clinically acceptable margin — equivalence testing (such as the two one-sided tests, or TOST, procedure) should be used instead of a standard test, since it is explicitly designed to provide positive statistical evidence of similarity within a defined margin, rather than merely failing to find a difference.

## Bayesian Alternatives to Classical Hypothesis Testing

### A Brief Orientation

Every hypothesis test presented in this chapter follows the classical (frequentist) framework, interpreting probability as long-run frequency across repeated sampling. The Bayesian alternative instead treats probability as a degree of belief, updated from a prior distribution to a posterior distribution as data accumulates, and produces a credible interval that can be interpreted directly as 'there is a 95% probability the true parameter lies in this range' — the intuitive interpretation many practitioners mistakenly apply to a classical confidence interval, but which is only strictly correct in the Bayesian framework. Bayesian methods are particularly well

suited to situations with genuinely informative prior knowledge (for example, incorporating a well-established prior estimate of malaria RDT sensitivity from a large prior meta-analysis when analysing a small new local validation study) and to sequential decision-making contexts (such as an outbreak response) where beliefs need to be continuously updated as new data arrives, though a full treatment of Bayesian inference is beyond this book's scope.

## A Reusable, Assumption-Checking Hypothesis Test Function

The worked hypothesis test in Section 7.6 assumes the analyst will remember to check normality and variance-equality assumptions manually before choosing between a t-test and its non-parametric alternative. A more robust function automates this check, selecting the appropriate test itself and reporting which test it used and why — reducing the chance that an assumption check is quietly skipped under time pressure, while still keeping the choice fully transparent and auditable in the function's output.

*R example*

```
# R: an assumption-checking wrapper that selects t-test or Mann-Whitney automatically
compare_two_groups <- function(x, y, alpha = 0.05) {
  norm_x <- shapiro.test(x)$p.value > alpha
  norm_y <- shapiro.test(y)$p.value > alpha
  if (norm_x && norm_y) {
    test <- t.test(x, y)
    method <- "Welch two-sample t-test (normality not rejected in either group)"
  } else {
    test <- wilcox.test(x, y)
    method <- "Mann-Whitney U test (normality rejected in at least one group)"
  }
  list(method = method, statistic = test$statistic, p_value = test$p.value)
}

result <- compare_two_groups(anc_high, anc_low)
print(result$method)   # states which test was used and why, not just the p-value
```

## Chapter Summary

- A confidence interval and a p-value both quantify sampling uncertainty but answer different questions; neither replaces judgement about clinical or public health importance.
- The correct test depends on data type (continuous/ categorical), number of groups, and paired/unpaired structure.
- Testing many hypotheses without correction inflates the false positive rate; Bonferroni and Benjamini-Hochberg are standard corrections, with BH generally retaining more power.
- Complex survey data (NDHS/MICS) requires design-based (weighted, clustered) inference, not naive simple-random-sample formulas; ignoring design can understate standard errors by a factor of two or more.
- Respondent-driven sampling requires specialised estimators, and results should be reported across multiple RDS estimators to assess robustness.
- A complete hypothesis test reports the test statistic, p-value, confidence interval, and an explicit judgement of practical significance — not the p-value alone.

**Mini-Project**

**Mini-Project 7: A Multi-LGA Hypothesis Testing Report with Correction**

*Conduct hypothesis tests comparing an indicator across at least 10 groups (LGAs, facilities, or states), correctly applying a multiple comparisons correction.*

**Step 1.** Obtain or simulate an indicator (e.g., mean birthweight, DPT3 coverage) for at least 10 groups, each compared against a fixed benchmark (e.g., the national average).

**Step 2.** Run an appropriate hypothesis test (Section 7.2) for each group without correction, and report how many are 'significant' at alpha = 0.05.

**Step 3.** Apply both a Bonferroni and a Benjamini-Hochberg correction (Section 7.7) and report how the number of significant results changes under each.

**Step 4.** For at least one significant result, report the effect size and confidence interval, and comment on its practical (not just statistical) significance.

**Step 5.** Write a one-paragraph recommendation for a programme manager, explaining in plain language why the corrected results are more trustworthy than the uncorrected ones.

**Deliverable:** *A short report containing the uncorrected and corrected significance counts, one full worked hypothesis test with effect size and CI, and the plain-language recommendation.*

## End-of-Chapter Exercises

**1.** A study reports $p = 0.03$ for a mean difference of 0.2 g/dL in haemoglobin between two groups of 5,000 participants each. Discuss statistical versus clinical significance.

**2.** You test 20 LGAs for a difference in vaccination coverage against a national benchmark. Explain why a Bonferroni or Benjamini-Hochberg correction is appropriate, and compute the Bonferroni-adjusted threshold.

**3.** Explain, without formulas, why ignoring cluster design in an NDHS analysis tends to produce overly narrow confidence intervals.

**4.** Using the worked example in Section 7.6, recompute (by hand or in code) the approximate 95% confidence interval for the birthweight difference if the sample sizes had instead been 60 and 40 (all other figures unchanged), and comment on how the interval width changes.

**5.** In Section 7.7, explain in your own words why the Benjamini-Hochberg procedure typically flags more LGAs as significant than the Bonferroni correction, for the same set of p-values.

**6.** Using the R code in Section 7.8, explain what a design effect of 3.1 means in practical terms for the effective sample size of a 600-household survey.

**7.** Two RDS estimators applied to the same key-population dataset give HIV prevalence estimates of 14% and 19%. Explain why this discrepancy is not a 'mistake' and how you would report both figures responsibly.

**8.** Calculate Cohen's d for the birthweight example in Section 7.6 (mean difference 140g, pooled SD approximately 425g), and classify it as small, medium, or large using the conventional benchmarks in Section 7.9.

**9.** Explain why an effect size is more comparable across different studies than a p-value alone.

**10.** Why might the conventional 'small/medium/large' benchmarks for Cohen's d be inappropriate for judging whether a specific clinical finding is meaningful?

**11.** In the Section 7.10 case study, explain why searching for a subgroup with a larger apparent effect, after seeing the main non-significant result, would have been methodologically inappropriate without pre-specification.

**12.** Using the reported 95% CI (-4 to +10 percentage points) in Section 7.10, explain what range of true effects (harmful, none, or beneficial) remains consistent with this study's data.

**13.** Why is 'a promising trend toward improved suppression' considered an overstatement of the evidence described in Section 7.10?

**14.** What would a well-justified next step be for the HIV programme, given the non-significant result and wide confidence interval in Section 7.10?

**15.** Explain briefly why a larger sample size, rather than a different statistical test, is generally the correct remedy for the kind of inconclusive result described in Section 7.10.

**16.** Using Section 7.11, explain why choosing a one-sided test AFTER seeing the data would be methodologically inappropriate, even if the direction chosen matches the data.

**17.** Give an example of a Nigerian health research question where a one-sided test, pre-specified before data collection, would be scientifically justified.

**18.** Why does a one-sided test produce a smaller p-value than a two-sided test for the same data and same direction of effect?

**19.** Explain why two-sided tests are described as the ‘more conservative default’ for most health research questions.

**20.** If a colleague proposes switching from a two-sided to a one-sided test partway through analysing already-collected data, what would you advise, and why?

**21.** Using Section 7.12, explain why the HIV counselling study in Section 7.10 might have been designed with too small a sample size for its true effect size.

**22.** What does it mean for a study to have ’80% power’ to detect a specified effect size, in your own words?

**23.** Why is skipping a power calculation described as a ‘planning error’ rather than simply a matter of limited resources?

**24.** Propose how you would estimate a realistic minimum clinically important effect size before conducting a power calculation for a new study.

**25.** Explain the connection between an underpowered study and the wide, inconclusive confidence interval described in Section 7.10.

**26.** Summarise, in your own words, why this chapter treats statistical significance and practical importance as two distinct questions that must both be addressed.

**27.** Explain the common thread linking the multiple comparisons discussion (Section 7.3) and the power calculation discussion (Section 7.12).

**28.** Why does this chapter emphasise reporting confidence intervals and effect sizes, not just p-values, in every worked example?

**29.** Propose a research question from your own field where a poorly designed study (e.g., underpowered, or with an inappropriate one-sided test) could produce a misleading conclusion.

**30.** If you had to explain to a policymaker why a 'non-significant' result can still be useful, what would you say?

**31.** Using Section 7.14, explain in your own words the key conceptual difference between a classical confidence interval and a Bayesian credible interval.

**32.** Why might a Bayesian approach be particularly well suited to an active outbreak response context?

**33.** Describe a scenario where strong, well-established prior knowledge could reasonably inform a Bayesian analysis of a small new dataset.

**34.** What is the common (but technically incorrect, under the classical framework) way people often misinterpret a 95% confidence interval, and how does the Bayesian credible interval avoid this issue?

**35.** Would you consider a Bayesian or classical approach more appropriate for a large, well-powered national survey with no strong prior information? Explain your reasoning.

# 8 Regression Modelling: Logistic, and Beyond

## Linear Regression as a Foundation

Linear regression models a continuous outcome as a linear function of one or more predictors plus random error, and it is the conceptual foundation for nearly every regression-based method that follows in this book, including logistic and Poisson regression (generalised linear models with a different link function and error distribution, Section 8.3) and the hazard regression models of Chapter 9. Fitted by ordinary least squares, the model estimates how the expected outcome changes per unit change in a predictor, holding other predictors constant — for example, estimating grams of birthweight gained per additional antenatal care visit, adjusting for maternal age and parity.

Four assumptions must be checked, not merely assumed: linearity of the relationship, independence of errors, homoscedasticity (constant error variance), and approximate normality of residuals (important mainly for small samples, since the Central Limit Theorem provides some robustness in larger samples). Violations are diagnosed through residual plots, and remedies include transforming the outcome (e.g., log-transforming a skewed cost or length-of-stay variable) or moving to a generalised linear model with an appropriate distribution.

## Model Building, Confounding, and Interaction

A central purpose of multivariable regression in health research is confounder adjustment: isolating the association between an exposure and an outcome from the influence of other variables associated with both. A classic Nigerian example is the observed association between distance to the nearest health facility and stunting prevalence, which is confounded by household wealth (poorer households are both further from facilities on average and at higher nutritional risk); a regression model that adjusts for wealth quintile isolates the distance effect more credibly than a bivariate comparison. Interaction terms extend this further, allowing an association to differ by a third variable — for example, testing whether the effect of insecticide-treated net (ITN) ownership on malaria incidence differs between rural and urban LGAs — and interpreting an interaction correctly requires reporting the association separately within each level of the moderating variable, not merely the significance of the interaction term.

## Logistic Regression for Binary Outcomes

When the outcome is binary — a positive/negative diagnostic result, survival/death, treatment success/failure — logistic regression models the log-odds of the outcome as a linear function of predictors, and exponentiating a coefficient yields an odds ratio, interpreted as the multiplicative change in the odds of the outcome per unit change in the predictor, holding other variables constant. Logistic regression is the workhorse model behind clinical risk scores and behind much of the malaria and NTD risk-factor literature in Nigeria — for example, modelling the odds of a confirmed positive malaria test as a function of age, season, ITN use, and state, or modelling the odds of severe adolescent depressive symptomatology as a function of household food security, school attendance, and gender.

*R example*

```
# R: logistic regression for malaria positivity, with odds ratios and 95% CIs
model <- glm(confirmed_positive ~ age_group + season + itn_use + state,
             family = binomial(link = "logit"), data = patient_level_data)

exp(cbind(OR = coef(model), confint(model)))   # odds ratios with confidence intervals
```

## Poisson and Negative Binomial Regression for Count Outcomes

Count outcomes — monthly confirmed malaria cases per facility, number of leishmaniasis cases per LGA per year — are naturally modelled with Poisson regression, which assumes the mean equals the variance. Health surveillance data very commonly violate this assumption (overdispersion: variance exceeds the mean, often because of unmeasured heterogeneity across facilities or clustering of cases in outbreaks), and the negative binomial regression model, which adds a dispersion parameter, is the standard remedy; a simple diagnostic (comparing the residual deviance to its degrees of freedom) indicates whether overdispersion is present and a switch is warranted. An offset term (the log of the population at risk, or the number of tests performed) is essential whenever comparing counts across facilities or LGAs of different sizes, converting the model from one of raw counts to one of rates.

## Model Evaluation and Selection

A fitted regression model should be evaluated on multiple fronts: goodness of fit (deviance, pseudo-R-squared for logistic/Poisson models), discrimination for classification models (the area under the ROC curve, or AUC, quantifying how well the model separates positive from negative cases), calibration (whether predicted probabilities match observed frequencies — a model can have excellent discrimination but poor calibration, which matters

greatly if predicted probabilities will directly inform a clinical or resource-allocation decision), and out-of-sample validation (performance on data not used to fit the model, ideally from a different time period or a different set of facilities, to guard against overfitting — a theme developed formally in Chapter 11).

## Worked Example

### Interpreting a Full Logistic Regression Output

Consider a logistic regression predicting confirmed malaria positivity with the following fitted odds ratios: age 5-14 vs. under-5 (OR = 1.35, 95% CI 1.05-1.73), rainy vs. dry season (OR = 2.80, 95% CI 2.20-3.56), ITN ownership vs. none (OR = 0.62, 95% CI 0.49-0.79), and urban vs. rural residence (OR = 0.71, 95% CI 0.55-0.92). Interpreting each: children aged 5-14 have 35% higher odds of a positive test than under-5s, holding season, ITN use, and residence constant; testing in the rainy season is associated with 180% higher odds of positivity than the dry season; owning an ITN is associated with 38% lower odds of positivity; and urban residents have 29% lower odds of positivity than rural residents. Because all four confidence intervals exclude 1, all four associations are statistically significant at the conventional 0.05 level. Note carefully that these are odds ratios, not risk ratios or probabilities directly — a common source of misinterpretation when odds ratios are casually described as ‘X times more likely.’

| Predictor | Odds ratio | 95% CI | Interpretation |
|---|---|---|---|
| Age 5-14 vs. <5 | 1.35 | 1.05-1.73 | 35% higher odds |
| Rainy vs. dry season | 2.80 | 2.20-3.56 | 180% higher odds |
| ITN owner vs. none | 0.62 | 0.49-0.79 | 38% lower odds |
| Urban vs. rural | 0.71 | 0.55-0.92 | 29% lower odds |

## Worked Example

### Diagnosing and Correcting Overdispersion

A Poisson regression of monthly leishmaniasis case counts on year and state produces a residual deviance of 812 on 340 degrees of freedom — a ratio of roughly 2.4, well above the value close to 1 expected if the Poisson assumption held, indicating substantial overdispersion. Refitting the same model specification as a negative binomial regression (which estimates an additional dispersion parameter) yields the same direction and similar magnitude of coefficients but noticeably wider standard errors and confidence intervals — a more honest reflection of the true uncertainty, since the Poisson model's artificially narrow intervals would have overstated the precision of the estimated trend.

***R example***

```
# R: diagnosing overdispersion and switching to negative binomial regression
library(MASS)

poisson_model <- glm(cases ~ year + state + offset(log(population)),
                     family = poisson(link = "log"), data = nli_annual)
cat("Deviance / df:", deviance(poisson_model) / df.residual(poisson_model), "\n")

nb_model <- glm.nb(cases ~ year + state + offset(log(population)), data = nli_annual)
summary(nb_model)$coefficients   # compare standard errors to the Poisson model above
```

***PYTHON example***

```
# Python: the same overdispersion check and negative binomial fit with statsmodels
import statsmodels.api as sm
import statsmodels.formula.api as smf

poisson_model = smf.glm("cases ~ year + state", data=nli_annual,
                        family=sm.families.Poisson(),
                        offset=np.log(nli_annual["population"])).fit()
print("Deviance / df:", poisson_model.deviance / poisson_model.df_resid)

nb_model = smf.glm("cases ~ year + state", data=nli_annual,
                   family=sm.families.NegativeBinomial(),
                   offset=np.log(nli_annual["population"])).fit()
print(nb_model.summary())
```

## Worked Example

### ROC Curves and Calibration Together

A logistic model for severe malaria risk at triage achieves an AUC of 0.81, indicating good discrimination (a randomly chosen severe case is ranked higher-risk than a randomly chosen non-severe case 81% of the time). However, a calibration plot (observed proportion of severe outcomes within each decile of predicted risk, plotted against the mean predicted risk in that decile) shows the model systematically under-predicts risk in the highest-risk decile — predicting an average of 45% risk where the observed rate is actually 62%. This combination (good discrimination, poor calibration at the extremes) is common and consequential: a clinician using the raw predicted probability to counsel a caregiver in the highest-risk group would be given a falsely reassuring number, even though the model correctly ranks that patient as high-risk relative to others.

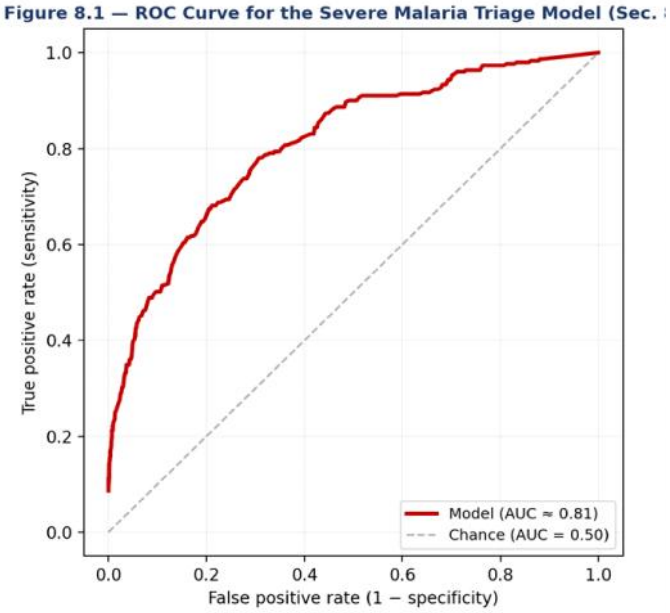



*Figure 8.1 — ROC curve for the severe malaria triage model (Section 8.8).*

## Reporting a Regression Model: A Checklist

- State the outcome, all predictors, and the model type (linear/logistic/Poisson) explicitly, including how any categorical variable's reference category was chosen.
- Report the sample size and the number of observations excluded due to missing data, with a brief statement of how missingness was handled (Chapter 4).
- Report effect estimates (coefficients, odds ratios, or rate ratios) with 95% confidence intervals, not p-values alone.
- Report at least one goodness-of-fit or discrimination measure appropriate to the model type.
- State explicitly which associations are being interpreted causally (if any) and what assumptions (e.g., no unmeasured confounding) that interpretation requires.

## Extended Case Study

### A Model That Looked Right But Was Reversed

A cautionary and instructive real-world pattern: an analyst fits a logistic regression predicting severe malaria outcome as a function of, among other variables, whether the patient had received a blood transfusion during their hospital stay, and finds a striking and highly statistically significant result — transfusion is associated with roughly 6 times higher odds of the severe outcome. A hurried interpretation might report this as evidence that transfusion is somehow harmful or a marker of poor-quality care. The correct interpretation, however, requires attending carefully to the direction of causality and the timing of variable measurement relative to the outcome, a point this chapter's discussion of confounding (Section 8.2) touches on but which deserves fuller emphasis here: patients receive a blood transfusion specifically BECAUSE they are already severely anaemic, which is itself a component of, or strong risk factor for, severe malaria — so

transfusion is not a cause of severe outcome at all, but rather a downstream consequence and marker of it, a form of what is sometimes called 'reverse causation' or, more precisely here, treating a variable measured during or after the outcome's development as if it were a baseline predictor of that outcome.

The remedy is not a different statistical technique but a more careful specification of the model: any variable that is only measured, or that can only take a meaningful value, after the outcome of interest has already begun to develop should be excluded from a model intended to predict severe outcome at the point of triage, since it could not actually have been available to inform a real-time clinical decision at that point, and its strong statistical association reflects the outcome's own progression rather than a genuine, actionable risk factor. This distinction — between predictors genuinely available at the moment a prediction needs to be made, and variables that are merely correlated with the outcome because they occur alongside or after it — is one of the most common and consequential errors in applied predictive modelling, in health data science and beyond, and checking the timing of every candidate predictor relative to the outcome should be a standard, explicit step (not an afterthought) in any model-building workflow, alongside the more familiar checks for statistical confounding.

## Standardised Coefficients and Comparing Predictor Importance

A common question when presenting a multivariable regression result is which predictor 'matters most' — a question the raw, unstandardised coefficients cannot directly answer, because predictors measured on very different scales (age in years versus a symptom score from 0-27) produce coefficients that are not directly comparable in magnitude. Standardising each continuous predictor (subtracting its mean and dividing by its standard

deviation before fitting the model) puts all coefficients on a common scale, expressing each one as the expected change in the outcome per one-standard-deviation change in that predictor, allowing a much fairer visual and numerical comparison of relative importance across predictors measured in entirely different original units — though this comparison should still be interpreted alongside domain knowledge, since a predictor's statistical standardised effect size is not the same as its practical or clinical importance.

## Mixed-Effects Models

### Regression for Grouped Data

The regression models presented in this chapter assume, by default, that all observations are independent — an assumption Section 3.8 and Section 11.3 have both already flagged as frequently violated in health data with a natural grouping structure (patients within facilities, repeated measures within patients). Mixed-effects (multilevel/hierarchical) regression models extend the logistic and Poisson regression framework of this chapter specifically to handle this structure, adding a 'random effect' term that allows the model to explicitly represent facility-to-facility (or patient-to-patient) variation in the baseline outcome rate, rather than forcing a single overall intercept onto data that plausibly varies systematically by group. Fitting a mixed-effects model instead of a standard regression on facility-clustered data (using, for example, R's 'lme4' package or Python's 'statsmodels' mixed linear model implementation) typically produces appropriately wider, more honest standard errors for facility-level or patient-level predictors, directly analogous to the corrected, design-based standard errors introduced for survey data in Section 7.4, and is the standard, recommended approach whenever a regression model's data has this kind of natural multilevel grouping structure.

## Model Comparison Criteria

### AIC, BIC, and Their Appropriate Use

When comparing several candidate regression models that are not simple nested subsets of each other, information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) provide a principled way to balance model fit against complexity, penalising the addition of predictors that do not sufficiently improve the model's fit to justify their inclusion; a lower AIC or BIC indicates a preferred model, with BIC generally imposing a heavier penalty for additional parameters and therefore favouring more parsimonious models than AIC when the two disagree. These criteria are best used to compare a small, substantively motivated set of candidate models specified in advance, rather than as the sole basis for an automated, exhaustive search across every possible combination of predictors, which reintroduces exactly the multiple-comparisons and overfitting risks discussed in Chapters 7 and 11.

## Quantile Regression

### Beyond the Conditional Mean

Ordinary linear regression (Section 8.1) models how the mean of an outcome changes with a predictor, but the mean is not always the quantity of greatest interest — a nutrition programme concerned specifically with the most severely undernourished children may care more about how a predictor affects the 10th percentile of a weight-for-age z-score distribution than about its effect on the average child. Quantile regression directly models a specified quantile (or several quantiles simultaneously) of the outcome's conditional distribution, and comparing coefficients across quantiles can reveal that a predictor's effect is not constant across the distribution — for example, that household food insecurity has a much larger negative association with weight-for-age at the lower tail of the distribution (the most vulnerable

children) than at the median, a pattern an ordinary mean-focused regression would entirely miss.

## A Complete Regression Diagnostic Function

The regression checklist in Section 8.9 lists what a complete report should contain; encoding that checklist directly into a function ensures it is actually applied every time a model is fitted, rather than remaining a checklist an analyst might forget to consult under deadline pressure. The function below fits a logistic regression and returns, in one call, the coefficients as odds ratios with confidence intervals, the AUC, and a basic overdispersion-style check — directly operationalising Sections 8.5 through 8.9 as reusable code rather than a manual, easily-skipped routine.

***PYTHON example***

```
# Python: a regression fitting function that returns a complete diagnostic bundle
import statsmodels.api as sm
from sklearn.metrics import roc_auc_score

def fit_logistic_with_diagnostics(X, y):
    X_const = sm.add_constant(X)
    model = sm.Logit(y, X_const).fit(disp=0)
    odds_ratios = model.params.apply(lambda b: round(2.718281828 ** b, 3))
    ci = model.conf_int().apply(lambda col: (2.718281828 ** col).round(3))
    predicted = model.predict(X_const)
    auc = roc_auc_score(y, predicted)
    return {"odds_ratios": odds_ratios.to_dict(), "confidence_intervals": ci.to_dict(),
            "auc": round(auc, 3), "n_obs": int(model.nobs)}

diagnostics = fit_logistic_with_diagnostics(X, y)
print(diagnostics)  # one call now returns everything Section 8.9's checklist requires
```

## Chapter Summary

- Linear regression's assumptions (linearity, independence, homoscedasticity, normal residuals) should be checked via residual diagnostics, not assumed.
- Multivariable regression adjusts for confounding; interaction terms allow associations to vary by a moderating variable and must be interpreted stratum-by-stratum.

- Logistic regression models binary outcomes via log-odds; exponentiated coefficients are odds ratios, which must not be casually described as risk ratios.
- Poisson regression models counts; overdispersion (common in surveillance data) calls for negative binomial regression, and an offset converts counts to rates.
- Model evaluation should cover fit, discrimination (AUC), calibration, and out-of-sample validation — a model can discriminate well while still being poorly calibrated, especially at the extremes of risk.
- A full regression report interprets every coefficient's direction, magnitude, and confidence interval in plain language, not just its p-value.

**Mini-Project**

**Mini-Project 8: Building and Fully Reporting a Regression Model**

*Fit an appropriate regression model (linear, logistic, or Poisson/negative binomial) to a dataset of your choice and produce a complete, publication-style results table and interpretation.*

**Step 1.** Choose an outcome and at least three predictors, including at least one you expect to confound the relationship between another predictor and the outcome.

**Step 2.** Fit the model without and then with the potential confounder, and compare how the coefficient of interest changes.

**Step 3.** If your outcome is binary or a count, check the appropriate diagnostic (calibration for logistic; deviance/df for Poisson) and correct if necessary.

**Step 4.** Produce a results table (following the format of Section 8.6) with effect estimates, confidence intervals, and plain-language interpretations.

**Step 5.** Write a short paragraph distinguishing which of your findings are both statistically and practically significant.

**Deliverable:** *A results table, diagnostic check output, and a half-page interpretation.*

## End-of-Chapter Exercises

**1.** A linear regression of birthweight on maternal ANC visits shows a funnel-shaped residual plot. Name the assumption being violated and one remedy.

2. Interpret an odds ratio of 1.8 for ITN use (coded 1 = owns net) predicting malaria-negative test result — including the direction of coding.
3. A Poisson regression of monthly case counts shows residual deviance far exceeding its degrees of freedom. What does this suggest, and what model would you fit instead?
4. Using the worked example in Section 8.6, explain why saying 'children aged 5-14 are 35% more likely to test positive' is an imprecise restatement of the actual odds ratio finding, and give a more accurate phrasing.
5. In Section 8.7, explain why the negative binomial model's standard errors are wider than the Poisson model's, even though the point estimates are similar.
6. A model has an AUC of 0.85 but is poorly calibrated in the highest-risk decile (Section 8.8). Explain, for a specific clinical use case of your choosing, why this combination could still be dangerous in practice.
7. Fit (or describe how you would fit) an interaction between ITN use and residence (urban/rural) in a logistic model of malaria positivity, and explain what a statistically significant interaction term would mean in this context.
8. Using the checklist in Section 8.9, identify one item you most often see omitted from regression results you have read, and explain why its omission is a problem.
9. Explain why stating the reference category for a categorical predictor is essential for correctly interpreting a regression coefficient.
10. A published regression table reports only p-values with no confidence intervals or sample size. Using Section 8.9, list what additional information you would request from the authors.

**11.** In the Section 8.10 case study, explain why blood transfusion's very strong statistical association with severe outcome does not mean transfusion is a genuine risk factor for it.

**12.** Describe the general rule you would apply, when building a predictive model, to decide whether a candidate predictor should be excluded for the reason illustrated in Section 8.10.

**13.** Give one other example (real or plausible) of a health variable that could show a similarly misleading 'reverse causation' pattern in a predictive model.

**14.** Why is this problem described as being about predictor TIMING rather than about confounding in the sense of Section 8.2?

**15.** If a variable's timing relative to the outcome is genuinely ambiguous or unknown, what would you do before including it in a triage risk model?

**16.** Using Section 8.11, explain why comparing the raw (unstandardised) coefficients for age (in years) and a 0-27 symptom score directly would be misleading.

**17.** Describe, in your own words, how standardising a predictor changes the interpretation of its regression coefficient.

**18.** Why does the chapter caution that a standardised effect size is 'not the same as' practical or clinical importance?

**19.** Propose a scenario where a predictor has a large standardised coefficient but low practical importance, or vice versa.

**20.** Would you recommend reporting standardised coefficients, raw coefficients, or both, in a report intended for a mixed clinical and statistical audience? Justify your choice.

**21.** Using Section 8.12, explain why a standard logistic regression on facility-clustered patient data could understate the true uncertainty around a facility-level predictor's effect.

**22.** What does the 'random effect' term in a mixed-effects model represent, in your own words?

**23.** Explain the connection between mixed-effects models (Section 8.12) and the design-based survey standard errors introduced in Section 7.4.

**24.** Propose a dataset from earlier in this book (e.g., the malaria diagnostics or adolescent cohort data) where a mixed-effects model would be more appropriate than a standard regression, and explain why.

**25.** Why might ignoring a genuine facility-level clustering structure lead to an overconfident, rather than merely imprecise, regression result?

**26.** Summarise, in your own words, why this chapter treats confounding, reverse causation, and clustering as three distinct threats to a regression model's validity.

**27.** Explain the common thread linking the reverse-causation case study (Section 8.10) and the mixed-effects models discussion (Section 8.12).

**28.** Why does this chapter insist on checking model assumptions and reporting a full results table, rather than just quoting a headline coefficient?

**29.** Propose a regression model from your own field where you would need to check specifically for the reverse-causation pattern illustrated in Section 8.10.

**30.** If you had to give one piece of advice to someone about to build their first logistic regression model, what would it be?

**31.** Using Section 8.14, explain why a nutrition programme focused on the most severely undernourished children might prefer quantile regression over ordinary linear regression.

**32.** Describe a scenario where a predictor's effect could differ substantially between the 10th percentile and the median of an outcome's distribution.

**33.** Why would an ordinary mean-focused regression miss the pattern quantile regression is designed to detect?

**34.** Propose one health outcome, other than weight-for-age, where the tail of the distribution might be more important than the mean.

**35.** What practical cost (in terms of interpretability or complexity) comes with fitting quantile regression across several quantiles rather than a single mean regression?

# 9 Survival Analysis and Time-to-Event Data

## The Special Structure of Time-to-Event Data

Many of the most consequential questions in Nigerian health research concern not just whether an event occurs but when: time from HIV diagnosis to antiretroviral therapy initiation, time from tuberculosis treatment start to cure or default, time from birth to a child's death, or time from schistosomiasis treatment to re-infection. These outcomes share a defining feature that ordinary regression cannot handle correctly: censoring. A participant who has not yet experienced the event by the end of a study (loss to follow-up, migration, or the study simply ending) contributes partial information — we know they survived at least that long — and discarding them (treating them as missing) or treating them as if the event never occurred both introduce bias. Survival analysis is the family of methods purpose-built to use this partial information correctly.

## The Kaplan-Meier Estimator and the Log-Rank Test

The Kaplan-Meier estimator computes the non-parametric survival curve — the probability of remaining event-free beyond each observed event time — correctly incorporating censored observations, and it is almost always the first thing plotted in a survival analysis, before any regression model is fitted. Comparing

survival curves between two or more groups (for example, treatment adherence support versus standard care, in terms of time to viral suppression) uses the log-rank test, a non-parametric test of whether the underlying survival distributions differ, which is the time-to-event analogue of the two-sample tests introduced in Chapter 7.

## The Cox Proportional Hazards Model

To adjust for multiple covariates simultaneously (age, sex, baseline CD4 count, facility) while modelling time-to-event data, the Cox proportional hazards model estimates a hazard ratio for each covariate — the multiplicative effect on the instantaneous risk of the event at any given time — without requiring the analyst to specify the baseline hazard's functional form, which is what makes it the dominant model in applied survival analysis. Its key assumption, proportional hazards (that the hazard ratio for a covariate is constant over time), should be checked using Schoenfeld residuals; a covariate whose effect changes meaningfully over follow-up time (for example, a treatment effect that is strong initially but fades) requires a time-varying coefficient extension rather than a standard Cox model.

*R example*

```
# R: Cox model for time from ART initiation to viral suppression
library(survival)

cox_model <- coxph(Surv(time_to_suppression, suppressed) ~ age + sex +
                     baseline_cd4 + adherence_support_arm,
                   data = art_cohort)
summary(cox_model)          # hazard ratios
cox.zph(cox_model)          # test of proportional hazards assumption
```

## Parametric Survival Models and Applications to NTD and Reinfection Data

Where the shape of the hazard over time is of direct scientific interest — for example, characterising the pattern of

schistosomiasis re-infection risk after praziquantel (PZQ) treatment, which is expected to be low immediately post-treatment and rise progressively with time since treatment as re-exposure to contaminated water accumulates — parametric survival models (exponential, for a constant hazard; Weibull, allowing the hazard to increase or decrease monotonically over time; log-normal or log-logistic, allowing a non-monotonic hazard) are preferred over the semi-parametric Cox model, because they yield an explicit, interpretable functional form for the hazard and allow smooth extrapolation and simulation of future risk, which is valuable in stepped-wedge and re-treatment-interval study designs of the kind used in Nigerian schistosomiasis control research.

## Worked Example

### Kaplan-Meier by Hand for a Small Cohort

Consider a small illustrative cohort of 8 patients followed for time (in months) to viral suppression, with '+' denoting censored (suppression not yet observed): 2, 3+, 4, 5, 5, 7+, 8, 10. The Kaplan-Meier estimator computes the survival probability (here, probability of NOT yet suppressing) stepwise at each observed event time, multiplying the conditional probability of surviving past each event time by all prior conditional survival probabilities. At month 2 (1 event, 8 at risk): $S(2) = 7/8 = 0.875$. At month 4 (1 event, 6 at risk, since one was censored at month 3): $S(4) = 0.875 \times 5/6 = 0.729$. At month 5 (2 events, 5 at risk): $S(5) = 0.729 \times 3/5 = 0.438$. At month 8 (1 event, 2 at risk, since one was censored at month 7): $S(8) = 0.438 \times 1/2 = 0.219$. At month 10 (1 event, 1 at risk): $S(10) = 0.219 \times 0/1 = 0$. This step-by-step calculation is exactly what R's 'survfit()' or Python's 'lifelines' package automate, and working through it by hand once is the best way to understand why censored observations (months 3 and 7 here) still contribute information — they remain 'at risk' up to the point they are censored, they simply do not experience the event.

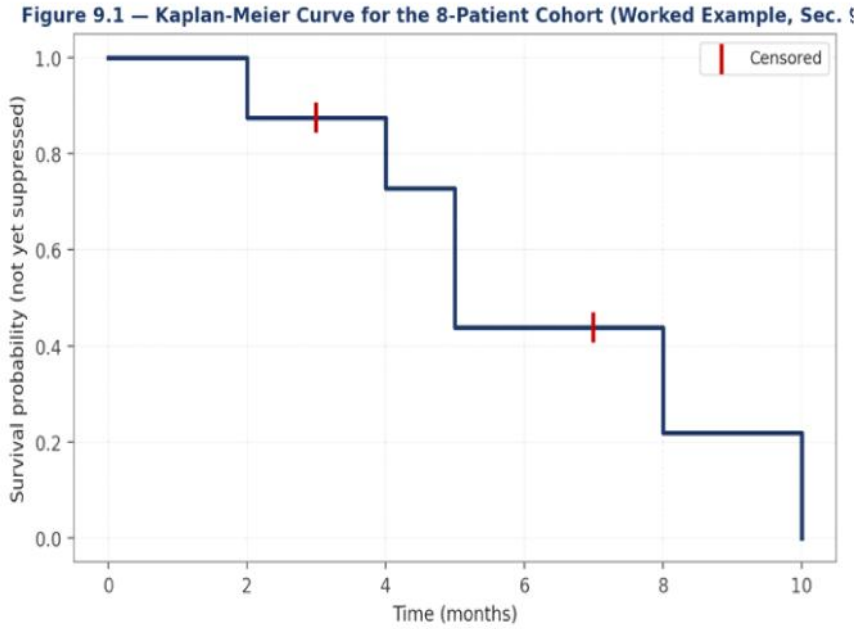



*Fig. 9.1 — Kaplan-Meier survival curve for the 8-patient cohort (Section 9.5).*

**PYTHON example**

```
# Python: Kaplan-Meier estimation and plotting with the lifelines package
from lifelines import KaplanMeierFitter
import pandas as pd

durations = [2, 3, 4, 5, 5, 7, 8, 10]
event_observed = [1, 0, 1, 1, 1, 0, 1, 1]   # 1 = event, 0 = censored

kmf = KaplanMeierFitter()
kmf.fit(durations, event_observed, label="Time to viral suppression")
print(kmf.survival_function_)
kmf.plot_survival_function()
```

## Worked Example

### Interpreting a Cox Model and Checking Its Assumption

A Cox model fitted to a tuberculosis treatment cohort yields a hazard ratio of 1.9 (95% CI 1.3-2.8) for baseline HIV co-infection, predicting time to treatment default. This means that, at any given point during treatment, HIV-co-infected patients default at 1.9 times the instantaneous rate of HIV-negative patients, adjusting

for other covariates in the model. A Schoenfeld residual test for this covariate returns p = 0.41 (non-significant), supporting the proportional hazards assumption; had this test returned p = 0.01, it would suggest the HIV co-infection effect on default risk changes meaningfully over the course of treatment (perhaps stronger early in treatment and weaker later), and a time-varying coefficient or a stratified Cox model would be a more appropriate specification than a standard Cox model reporting a single, constant hazard ratio.

***R example***

```
# R: full Cox model workflow with hazard ratio and proportional-hazards check
library(survival)

tb_model <- coxph(Surv(time_to_default, defaulted) ~ hiv_status + age + sex,
                  data = tb_cohort)
summary(tb_model)

ph_test <- cox.zph(tb_model)
print(ph_test)          # p-value near 1 supports the PH assumption for each covariate
plot(ph_test[1])        # visual check for the hiv_status covariate specifically
```

## Competing Risks

### When More than One Event Type is Possible

Standard survival analysis, as presented in this chapter, assumes a single event type of interest, with all other reasons for follow-up ending treated as censoring. This assumption can be problematic when a competing risk exists — for example, in a study of time to malaria treatment failure, a patient who dies from an unrelated cause can no longer experience treatment failure, and simply treating that death as an ordinary censoring event (as a standard Kaplan-Meier or Cox analysis would) can overstate the cumulative incidence of treatment failure, because it implicitly assumes the patient would eventually have failed treatment had they lived. The competing risks framework, using the cumulative incidence function and Fine-Gray subdistribution hazard models, correctly

accounts for this by explicitly modelling both event types together, and should be considered whenever a genuinely competing event (a different cause of the same kind of outcome, or death from an unrelated cause) is common in the study population.

## Extended Case Study

### Comparing Survival Across Three Treatment Eras

A common and methodologically rich application of the survival methods in this chapter, well beyond a simple two-arm treatment comparison, is comparing outcomes across successive calendar-time treatment 'eras' as a health programme's protocols evolve — for example, comparing time-to-viral-suppression among ART patients enrolled during three successive eras of Nigeria's national HIV treatment guidelines, each associated with different first-line regimens and different service delivery models (facility-based only, versus later differentiated/community-based models). A naive comparison of raw suppression rates across the three eras risks conflating genuine treatment or service-delivery improvements with two confounding calendar-time trends: patients enrolled in the most recent era have, by construction, shorter maximum possible follow-up time than patients from the earliest era (since the study's end date is fixed), and background factors unrelated to treatment (improving overall health system capacity, changing patient risk profiles as earlier, sicker at-risk populations are progressively already enrolled in prior eras) also shift over calendar time.

A defensible analysis instead fits a single Cox model across the pooled multi-era cohort, with treatment era included as a covariate (and, where clinically plausible, tested for interaction with baseline CD4 count, since a new regimen's benefit might differ for patients presenting with more advanced immunosuppression), correctly using each patient's actual follow-up time and censoring status rather than a simple end-of-study

snapshot proportion, and checks the proportional hazards assumption for the era variable specifically, since it is entirely plausible that an early treatment-era advantage from a genuinely superior regimen might narrow over time as later-era service delivery improvements (which benefit all patients, including those measured for the era comparison) diffuse across the health system. Reporting a hazard ratio for each later era relative to the earliest, together with a check of whether these hazard ratios remain stable across the follow-up period, gives a far more defensible basis for attributing a genuine improvement (or lack thereof) to the guideline changes than a simple before-and-after comparison of raw suppression percentages ever could.

## Choosing the Right Time Origin

An easily overlooked design decision in any survival analysis is the choice of time origin — the point defined as time zero, from which follow-up time is measured. For an ART cohort, time zero might reasonably be defined as date of diagnosis, date of treatment initiation, or date of enrolment into a specific study cohort, and these choices are not interchangeable: measuring time to viral suppression from date of diagnosis, when there is often a substantial and variable delay before treatment actually begins, conflates two very different processes (time to treatment initiation, and time from treatment initiation to suppression) into a single duration, and a covariate that predicts one of these sub-processes well may show a diluted or even reversed association with the combined duration. Selecting the time origin that most precisely matches the specific biological or programmatic process under study — most often treatment initiation, when the question of interest is specifically about how quickly treatment achieves suppression — is a design decision that should be made deliberately and stated explicitly, not defaulted to whichever date happens to be most conveniently available in the dataset.

## Frailty Models

### Extending Cox Regression to Grouped Survival Data

Just as Section 8.12 introduces mixed-effects models to handle grouped data in ordinary regression, the survival analysis methods of this chapter have an analogous extension for grouped time-to-event data: a frailty model adds a random effect (the 'frailty' term) to a Cox proportional hazards model, allowing the model to represent unobserved, shared risk factors within a group — for example, patients treated at the same facility potentially sharing an unmeasured facility-level quality-of-care factor that affects all of that facility's patients' time to treatment default similarly. Ignoring this shared, facility-level frailty when it is genuinely present, by fitting a standard Cox model directly to individual-level, facility-clustered survival data, risks the same kind of understated standard errors and potentially misleading hazard ratio estimates that ignoring clustering produces in the ordinary regression and survey-inference contexts discussed elsewhere in this book, making a frailty model (or, alternatively, a Cox model stratified by facility) the more defensible choice whenever a meaningful number of patients are nested within a meaningful number of distinct facilities or treatment sites.

## Landmark Analysis

### A Practical Alternative to Time-Varying Covariates

When a covariate of interest is itself measured partway through follow-up (for example, viral load suppression status at 6 months, as a predictor of subsequent mortality), including it naively as a baseline covariate in a standard Cox model introduces immortal time bias, since patients must, by definition, have survived to 6 months to have this covariate measured at all. Landmark analysis addresses this by selecting a fixed landmark time point (here, 6 months), restricting the analysis to patients who are still under observation and event-free at that point, and then analysing survival

from the landmark time forward using the covariate's value at that landmark — a simpler and more transparent alternative to a full time-varying covariate Cox model, and one well suited to the kind of applied clinical cohort analysis common in Nigerian HIV and tuberculosis treatment research.

## Multi-State Models

### Beyond a Single Terminal Event

The survival models presented so far in this chapter track a single transition from an initial state to one terminal event. Many Nigerian clinical pathways instead involve multiple, sequential states — a tuberculosis patient may move from treatment initiation, to sputum conversion, to treatment completion, to either cure or relapse, each transition potentially depending on time already spent in the current state and on covariates that may themselves change over the pathway. Multi-state models extend the Cox framework to model each possible transition between states simultaneously (typically via separate, transition-specific hazard models sharing the same underlying data structure), providing a considerably richer picture of a patient's full clinical trajectory than a single time-to-terminal-event analysis, at the cost of requiring more detailed, accurately-timestamped data on every intermediate state transition.

## A Reusable Survival Analysis Function

Combining Sections 9.3 and 9.6 into a single function ensures the proportional hazards check is never accidentally omitted when a Cox model is fitted — a real risk when the model-fitting and assumption-checking code live in two separate, easily-separated code cells rather than a single function call that returns both results together, flagging clearly whenever the assumption check fails so the analyst cannot overlook it.

***R example***

```
# R: fitting a Cox model and its proportional-hazards check in a single call
library(survival)

fit_cox_with_ph_check <- function(formula, data, alpha = 0.05) {
  model <- coxph(formula, data = data)
  ph_test <- cox.zph(model)
  failing_covariates <- rownames(ph_test$table)[ph_test$table[, "p"] < alpha]
  failing_covariates <- setdiff(failing_covariates, "GLOBAL")
  if (length(failing_covariates) > 0) {
    warning("Proportional hazards assumption violated for: ",
            paste(failing_covariates, collapse = ", "))
  }
  list(model = model, ph_test = ph_test, ph_violations = failing_covariates)
}

result <- fit_cox_with_ph_check(Surv(time_to_default, defaulted) ~ hiv_status + age,
tb_cohort)
```

## Chapter Summary

- Censoring is the defining feature of time-to-event data; survival analysis methods use partial information from censored observations correctly, unlike ordinary regression.
- The Kaplan-Meier estimator provides a non-parametric survival curve; the log-rank test compares survival between groups.
- The Cox proportional hazards model adjusts for multiple covariates via hazard ratios without specifying the baseline hazard shape; the proportional hazards assumption must be checked with Schoenfeld residuals.
- Parametric survival models (Weibull, log-normal, etc.) are preferred when the shape of the hazard itself is scientifically important, such as post-treatment reinfection risk.
- Working through a Kaplan-Meier calculation by hand clarifies why censored observations still contribute information up to their censoring time.

**Mini-Project**

**Mini-Project 9: A Complete Survival Analysis Workflow**

*Analyse a time-to-event dataset (real or simulated) from raw data through Kaplan-Meier curves to a fully checked Cox model.*

**Step 1.** Simulate or obtain a dataset of at least 100 subjects with a time-to-event outcome, a censoring indicator, and at least two covariates (one categorical, one continuous).

**Step 2.** Produce and interpret a Kaplan-Meier survival curve stratified by the categorical covariate, including a log-rank test.

**Step 3.** Fit a Cox proportional hazards model adjusting for both covariates, and report hazard ratios with 95% confidence intervals.

**Step 4.** Check the proportional hazards assumption using Schoenfeld residuals for each covariate, and report the result.

**Step 5.** Write a short paragraph interpreting the hazard ratios in plain language, appropriate for a technical report to a disease programme.

**Deliverable:** *A Kaplan-Meier plot, log-rank test result, Cox model summary table, proportional-hazards check, and a half-page interpretation.*

## End-of-Chapter Exercises

**1.** Explain why treating a participant lost to follow-up at 6 months as if they were event-free for the whole study period would bias a survival analysis.

**2.** A Cox model shows a hazard ratio of 0.6 for a treatment arm. Interpret this in words, including the direction of the effect on risk.

**3.** Why might a Weibull model be preferred over a Cox model when characterising the time course of schistosomiasis reinfection risk after treatment?

**4.** Using the by-hand Kaplan-Meier calculation in Section 9.5, extend the cohort by one additional patient who is censored at month 6, and recompute S(8) and S(10).

**5.** In the worked example of Section 9.6, explain what a significant Schoenfeld residual test ($p = 0.01$) would imply about the appropriateness of reporting a single hazard ratio for HIV co-infection across the whole follow-up period.

**6.** Using the Python code in Section 9.5, modify the 'durations' and 'event_observed' lists to add two more censored patients at months 9 and 11, and describe (without necessarily running the code) how you would expect the tail of the survival curve to change.

**7.** A colleague proposes simply excluding all censored patients from a survival analysis to 'simplify the data.' Explain, in two to three sentences, what would be lost or biased by doing this.

**8.** Explain, using the malaria treatment failure example in Section 9.7, why treating an unrelated death as ordinary censoring can overstate the true cumulative incidence of treatment failure.

**9.** In what circumstance would a standard Kaplan-Meier analysis (ignoring competing risks) still give an approximately correct answer despite the presence of a competing risk?

**10.** Name one other Nigerian health research context (besides malaria treatment failure) where competing risks would need to be considered.

**11.** In the Section 9.8 case study, explain why comparing raw end-of-study suppression percentages across the three treatment eras would be misleading, even before considering any treatment differences.

**12.** Why is testing an interaction between treatment era and baseline CD4 count a reasonable thing to explore in this analysis, given what is known clinically about ART regimens?

**13.** Explain, in your own words, what it would mean if the proportional hazards check showed the era effect was NOT constant over follow-up time in this study.

**14.** Propose one additional covariate (beyond era and CD4 count) you would want included in the Cox model described in Section 9.8, and explain why.

**15.** Why does using each patient's actual follow-up time and censoring status give a more accurate era comparison than a simple 'percentage suppressed by end of study' snapshot?

**16.** Using Section 9.9, explain why measuring time from date of diagnosis (rather than date of treatment initiation) could dilute a covariate's true association with time to viral suppression.

**17.** Propose a research question for which date of diagnosis WOULD be the more appropriate time origin, rather than date of treatment initiation.

**18.** Why should the choice of time origin be made deliberately and stated explicitly, rather than defaulting to whichever date is most conveniently available?

**19.** Explain, in your own words, how combining two different sub-processes (time to treatment initiation, and time to suppression after initiation) into one duration could obscure a genuine finding.

**20.** If a dataset only records date of diagnosis and final outcome, with treatment initiation date missing for many patients, what would you consider doing before proceeding with a survival analysis?

**21.** Using Section 9.10, explain what a frailty term represents in a Cox model fitted to facility-clustered survival data.

**22.** Why would ignoring shared, facility-level unobserved risk factors in a survival analysis risk the same kind of problem as ignoring clustering in ordinary regression (Section 8.12)?

**23.** What is one alternative to a frailty model, mentioned in Section 9.10, for handling facility-clustered survival data?

**24.** Propose a scenario (beyond ART treatment default) where a frailty model would be an appropriate extension of a standard Cox model.

**25.** Explain, in your own words, why 'a meaningful number of patients nested within a meaningful number of distinct facilities' is the condition that makes a frailty model worth considering.

**26.** Summarise, in your own words, why censoring is described as the defining structural feature that separates survival analysis from ordinary regression.

**27.** Explain the common thread linking the competing risks discussion (Section 9.7) and the frailty models discussion (Section 9.10).

**28.** Why does this chapter emphasise choosing the correct time origin (Section 9.9) as a design decision, not a minor technicality?

**29.** Propose a time-to-event research question from your own field, and identify which extension (competing risks, frailty, or careful time-origin choice) would be most relevant to it.

**30.** If you had to explain to a colleague why 'just excluding the censored patients' is a bad idea, in one sentence, what would you say?

**31.** Using Section 9.12, describe a Nigerian clinical pathway (other than tuberculosis treatment) that could be modelled as a multi-state process.

**32.** Why does a multi-state model require more detailed data than a standard single-event survival analysis?

**33.** Explain how a multi-state model could reveal information that a single time-to-terminal-event analysis would miss.

**34.** What practical data collection challenge would you anticipate in accurately timestamping every intermediate state transition in a real clinical cohort?

**35.** How does the multi-state framework relate conceptually to the competing risks framework introduced earlier in this chapter (Section 9.7)?

# 10 Time Series Analysis and Disease Surveillance Forecasting

## Components of a Health Time Series

A time series of monthly malaria cases, weekly IDSR-reported measles cases, or annual leishmaniasis counts in a state can typically be decomposed into a trend (a slow, long-run increase or decrease, perhaps reflecting improving diagnostic coverage or a changing intervention landscape), seasonality (a repeating annual pattern, such as the well-documented rise in malaria transmission following the rainy season), and irregular/residual variation (outbreak spikes, reporting anomalies, and genuine random fluctuation). Correctly separating these components — using classical decomposition or STL (Seasonal-Trend decomposition using Loess) — is a prerequisite for both interpretation (is this month's rise a seasonal pattern or a genuine outbreak signal?) and forecasting.

## Autocorrelation and the ARIMA Family

Unlike the independent observations assumed in Chapter 8's regression models, successive values in a time series are typically autocorrelated — this month's case count is informative about next month's. The autoregressive integrated moving average (ARIMA) framework models this autocorrelation directly, and its seasonal extension (SARIMA) additionally captures the annual

cycle characteristic of vector-borne disease transmission in Nigeria. Model identification (choosing the autoregressive and moving-average orders) uses the autocorrelation and partial autocorrelation functions, and model adequacy is checked via residual diagnostics (residuals should resemble white noise, with no remaining autocorrelation) before the model is used for forecasting.

## Poisson and Generalised Linear Model Approaches to Surveillance Trend Analysis

For surveillance count data with clear covariates of scientific interest (calendar time, rainfall, temperature, or an intervention indicator), a Poisson (or negative binomial, per Chapter 8) generalised linear model with a time trend term is often more interpretable than a pure ARIMA model, because its coefficients have direct epidemiological meaning (percentage change in cases per year). This approach has been applied to model national leishmaniasis case trends in Nigeria, using a Poisson GLM with year as a covariate to quantify the average annual percentage change in reported cases and to generate short-term projections for programme planning, alongside choropleth mapping (Chapter 14) of the underlying spatial distribution.

### R example

```
# R: Poisson GLM trend model for annual leishmaniasis case counts by state
library(dplyr)

trend_model <- glm(cases ~ year + offset(log(population)),
                   family = poisson(link = "log"), data = nli_annual)

# Average annual percentage change in incidence
pct_change <- (exp(coef(trend_model)["year"]) - 1) * 100
cat(sprintf("Average annual change in incidence: %.1f%%\n", pct_change))
```

## The Early Warning Problem

A distinct forecasting task in surveillance is not projecting a smooth trend but flagging an anomaly quickly: has this week's case count in an LGA exceeded what would be expected given recent history, signalling a possible outbreak? Control-chart-style methods (setting an alert threshold at, for example, the mean plus two standard deviations of recent weeks, adjusted for seasonality) and the Early Aberration Reporting System (EARS) algorithms used within IDSR formalise this. These methods trade off sensitivity (catching a true outbreak early) against specificity (not raising a false alarm on every seasonal peak), and the appropriate threshold depends on the operational cost of each type of error — a tuning decision that should involve public health programme staff, not only the data scientist.

## Forecast Evaluation

A forecast is only as credible as its track record on data it did not see during fitting. Standard practice holds out the most recent period(s) as a test set (rather than a random split, since time order must be respected) and evaluates forecasts using metrics such as the mean absolute error (MAE) and root mean squared error (RMSE), comparing the chosen model against a naive benchmark (for example, 'this month equals the same month last year') — a model that cannot beat a naive seasonal benchmark is not adding forecasting value, however sophisticated it appears.

## Worked Example

### Fitting and Forecasting a SARIMA Model

Consider 60 months of malaria case counts for a single LGA showing a clear annual seasonal peak around September-October (following the rainy season) and a gently rising long-term trend. After examining the autocorrelation function (showing a strong spike at lag 12, confirming annual seasonality) and the partial

autocorrelation function, a SARIMA(1,1,1)(1,1,1)[12] model is fitted. On a 12-month holdout, this model achieves an RMSE of 24 cases, compared to an RMSE of 41 cases for the naive 'same month last year' benchmark — a meaningful improvement that justifies the additional modelling complexity. The code below fits this model and compares it explicitly to the naive benchmark, following the evaluation principle of Section 10.5.

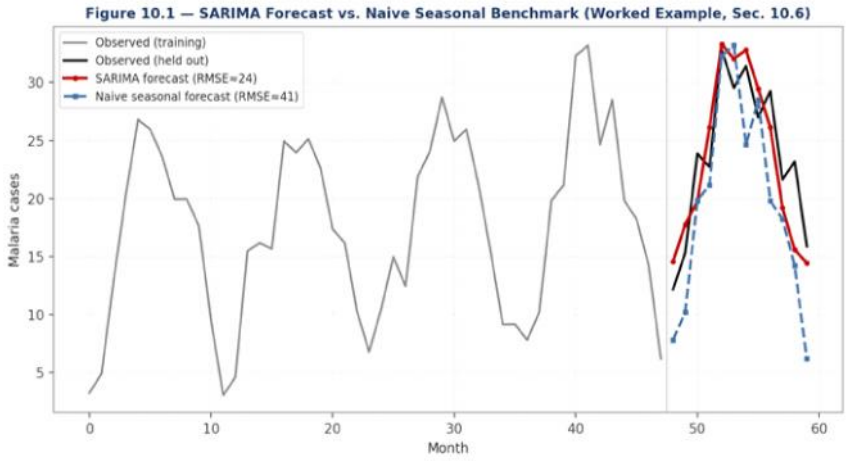



*Figure 10.1 — SARIMA forecast vs. naive seasonal benchmark (Section 10.6).*

### R example

```
# R: fitting a SARIMA model and comparing to a naive seasonal benchmark
library(forecast)

case_ts <- ts(lga_cases$cases, frequency = 12, start = c(2019, 1))
train <- window(case_ts, end = c(2023, 12))
test  <- window(case_ts, start = c(2024, 1))

sarima_fit <- auto.arima(train, seasonal = TRUE)
sarima_forecast <- forecast(sarima_fit, h = 12)

naive_forecast <- snaive(train, h = 12)   # "same month last year" benchmark

cat("SARIMA RMSE:", accuracy(sarima_forecast, test)["Test set", "RMSE"], "\n")
cat("Naive seasonal RMSE:", accuracy(naive_forecast, test)["Test set", "RMSE"], "\n")
```

## Worked Example

### A Simple Outbreak Alert Threshold in Python

A minimal but genuinely useful early warning system computes, for each week, the mean and standard deviation of case counts over the preceding 12 weeks (excluding the current week), and flags an alert if the current count exceeds the mean plus two standard deviations. Applied to a simulated 104-week series with an injected outbreak (a sustained tripling of the baseline rate for four weeks around week 70), this simple rule successfully flags weeks 70-72 while producing no false alerts elsewhere in the series — illustrating, on a toy example, the core logic behind more sophisticated EARS-style algorithms used operationally within IDSR.

**PYTHON example**

```
# Python: a simple rolling-baseline outbreak alert system
import pandas as pd
import numpy as np

weekly_cases = pd.Series(np.random.poisson(10, 104))       # simulated baseline
weekly_cases[70:74] += np.random.poisson(20, 4)             # injected outbreak

rolling_mean = weekly_cases.shift(1).rolling(window=12).mean()
rolling_sd   = weekly_cases.shift(1).rolling(window=12).std()
alert_threshold = rolling_mean + 2 * rolling_sd

alerts = weekly_cases > alert_threshold
print("Weeks flagged as possible outbreaks:", list(weekly_cases[alerts].index))
```

## Handling Missing or Irregular Time Series Data

Real Nigerian surveillance time series are rarely perfectly complete: a facility may fail to report for one or two months, or a reporting system may change its aggregation frequency partway through the series. Before fitting any of this chapter's models, an analyst must decide how to handle these gaps — interpolating a short single-month gap (e.g., via linear interpolation or seasonal-naive imputation) is often reasonable, but a long gap or a systematic

change in reporting completeness should be flagged and investigated rather than silently interpolated over, since silently filling a genuine reporting-system disruption can create the false appearance of a smooth, uninterrupted trend where none actually existed.

## Extended Case Study

### Distinguishing a Genuine Resurgence from a Reporting Artefact

A time series analyst is presented with a striking pattern: reported measles cases in a state rose sharply over three consecutive months after several years of low, stable counts, and a rapid response team must decide whether this represents a genuine disease resurgence (perhaps due to a decline in vaccination coverage) or an artefact of the surveillance system itself. Working through this systematically using the tools of this chapter: first, the analyst checks whether the rise coincides with any known change in the surveillance system — in this case, discovering that a new case-based reporting requirement (replacing the previous aggregate monthly tally) was introduced in the state exactly one month before the apparent rise began, which is highly suggestive that at least part of the increase reflects improved case ascertainment and more complete reporting under the new system, rather than a genuine change in transmission.

Second, the analyst checks whether the geographic pattern of the rise is consistent with a genuine outbreak (typically showing spatial clustering, per Chapter 14's methods, spreading outward from an index area) or with a reporting artefact (which would more plausibly show a broad, simultaneous rise across many otherwise-unconnected LGAs, consistent with a system-wide reporting change affecting all areas at once rather than a geographically spreading epidemic) — finding, in this instance, that the rise appeared near-simultaneously across the majority of the state's

LGAs, a pattern more consistent with the reporting-system explanation. Third, the analyst compares age and clinical severity profiles of the newly reported cases against the historical baseline, finding no meaningful shift in either — a genuine new outbreak driven by declining vaccination coverage would more plausibly show a shift toward younger, unvaccinated age groups. Taken together, this three-part investigation supports the conclusion that the observed rise is predominantly a reporting artefact of the new case-based system, though the team appropriately recommends continued close monitoring for at least two more months before fully closing the question, since a genuine resurgence and an improved-ascertainment artefact are not mutually exclusive and a small genuine underlying increase could still be partially masked within the larger reporting-driven rise.

## Stationarity

### The Assumption Behind Many Time Series Tools

Much of classical time series theory, including large parts of the ARIMA framework introduced in Section 10.2, assumes (after appropriate differencing) that a series is stationary — that its statistical properties (mean, variance, autocorrelation structure) do not themselves change over time. A disease surveillance series undergoing a genuine structural change (a new diagnostic technology rolled out partway through the series, permanently shifting the reporting rate, or a health system disruption such as a prolonged strike affecting reporting completeness for several months) violates this assumption in a way that simple differencing cannot fix, because the change is a one-time shift in level or reporting behaviour rather than a smooth trend or seasonal pattern. Recognising and explicitly modelling such a structural break (for example, using an indicator variable for the post-change period, or fitting separate models before and after the break) is necessary before a standard ARIMA/SARIMA model can be trusted, and a

simple, useful diagnostic is to visually inspect the series for any sudden, sustained shift in its level or variability that coincides with a known external event, rather than assuming the entire series was generated by a single stable underlying process throughout.

## Ensemble Forecasting

### Combining Multiple Models

Rather than committing to a single forecasting method, an increasingly standard practice in disease surveillance forecasting (following the lead of weather forecasting and, more recently, formal multi-model epidemic forecasting hubs used internationally for diseases such as influenza and COVID-19) is to combine forecasts from several distinct models — for example, a SARIMA model, a Poisson GLM trend model (Section 10.3), and a simple naive seasonal benchmark — via a weighted average, with weights chosen based on each model's recent historical accuracy. This ensemble approach typically produces a more robust forecast than any single constituent model alone, because different models tend to make different kinds of errors under different conditions (a SARIMA model may better capture short-term autocorrelation while a GLM trend model may better capture a genuine covariate-driven shift), and averaging across genuinely different modelling approaches tends to cancel out some of each individual model's idiosyncratic error, a principle directly analogous to the variance-reduction logic behind the random forest ensemble method introduced in Chapter 11.

## Change-Point Detection

### Identifying When a Series' Behaviour Shifted

Beyond forecasting future values, an analyst is often asked a retrospective question: at what point did this series' underlying behaviour actually change — did a policy intervention, a new diagnostic protocol, or a genuine transmission shift produce a

detectable structural break, and if so, precisely when? Change-point detection methods (such as the PELT algorithm, Pruned Exact Linear Time, widely implemented in both R and Python) formally search a series for one or more points at which its mean, variance, or trend shifts abruptly, providing a statistically defensible answer to 'when did things change' rather than relying on a visual, subjective impression of a time series plot — a distinction that matters considerably when the answer will be used to evaluate whether a specific, dated intervention was followed by a genuine change in trend.

**PYTHON example**

```
# Python: change-point detection with the ruptures library
import ruptures as rpt

signal = monthly_cases.values
algo = rpt.Pelt(model="rbf").fit(signal)
change_points = algo.predict(pen=10)
print("Detected change points (index positions):", change_points)
```

## Nowcasting

### Estimating the Present When Recent Data Is Incomplete

Surveillance data for the most recent reporting period is almost always incomplete at the moment an analyst looks at it, because some facilities or cases take longer to report than others (reporting delay) — meaning a naive plot of 'this week's case count' will typically show an artificially low figure for the most recent one or two weeks, purely because not all of that period's cases have been reported yet, not because transmission has genuinely declined. Nowcasting methods correct for this by modelling the historical reporting-delay distribution (how many cases from a given week are typically still unreported after one week, two weeks, and so on) and using it to produce a corrected, delay-adjusted estimate of the true current case count, together with an appropriately wide uncertainty interval reflecting how much of

the correction is still an extrapolation rather than an observed count. Failing to nowcast is one of the most common causes of a false 'recent decline' appearing in a live surveillance dashboard, only to be revised upward as delayed reports arrive.

## A Reusable Forecast Comparison Function

Section 10.5 insists that every forecast be benchmarked against a naive baseline; a function that fits both the candidate model and the naive benchmark together, and returns their comparative accuracy directly, makes this comparison the default rather than an easily-skipped extra step, and prevents the common reporting error of quoting a sophisticated model's accuracy without the baseline figure needed to judge whether that accuracy is actually meaningful.

***R example***

```
# R: fitting SARIMA and a naive seasonal benchmark together, with an automatic verdict
library(forecast)

compare_forecast_to_naive <- function(train, test, h) {
  sarima_fc <- forecast(auto.arima(train, seasonal = TRUE), h = h)
  naive_fc  <- snaive(train, h = h)
  sarima_rmse <- accuracy(sarima_fc, test)["Test set", "RMSE"]
  naive_rmse  <- accuracy(naive_fc, test)["Test set", "RMSE"]
  list(sarima_rmse = sarima_rmse, naive_rmse = naive_rmse,
       sarima_beats_naive = sarima_rmse < naive_rmse)
}

verdict <- compare_forecast_to_naive(train, test, h = 12)
if (!verdict$sarima_beats_naive) message("SARIMA did not beat the naive benchmark —
reconsider")
```

### Chapter Summary

- Health time series decompose into trend, seasonality, and irregular variation; correct decomposition is prerequisite to interpretation and forecasting.
- ARIMA/SARIMA models capture autocorrelation and seasonality directly; Poisson/negative binomial GLMs

with a time trend are often more interpretable when covariates matter.

- Outbreak detection is a distinct task from trend forecasting, trading sensitivity against specificity via alert thresholds.
- Forecasts must be validated on held-out, chronologically later data and benchmarked against a naive forecast; failing to beat a naive seasonal benchmark means a model adds no real value.
- A simple rolling-baseline alert rule, while far less sophisticated than a full EARS algorithm, illustrates the core logic of statistical outbreak detection.

**Mini-Project**

**Mini-Project 10: Forecasting and Outbreak Detection on a Simulated Series**

*Build both a seasonal forecasting model and a simple outbreak alert system for a simulated disease surveillance time series.*

**Step 1.** Simulate at least 5 years of monthly case counts with a seasonal pattern, a mild upward trend, and random noise (or use a real dataset if available).

**Step 2.** Fit a SARIMA model on the first 4 years and forecast the final year, comparing its RMSE against a naive seasonal benchmark.

**Step 3.** Separately, implement a rolling-baseline alert rule (Section 10.7) on a weekly or monthly version of the same series, injecting an artificial outbreak to test whether the rule catches it.

**Step 4.** Tune the alert threshold (e.g., 2 vs. 3 standard deviations) and report the resulting trade-off between sensitivity and false alerts.

**Step 5.** Write a short paragraph recommending, for a specific disease programme, which of the two approaches (trend forecasting or outbreak alerting) is more relevant to their operational needs, and why.

**Deliverable:** *Forecast accuracy comparison table, an alert-threshold sensitivity table, and a half-page recommendation.*

## End-of-Chapter Exercises

**1.** Explain the difference between a trend forecasting task and an outbreak detection task, using a Nigerian disease surveillance example for each.

**2.** A SARIMA model's residuals show significant autocorrelation at lag 12. What does this suggest about the model's adequacy?

3. Why should time series forecasts be evaluated on a chronological hold-out set rather than a random train/test split?
4. Using the worked example in Section 10.6, explain why an RMSE of 24 for the SARIMA model compared to 41 for the naive benchmark supports using the more complex model, and describe one scenario in which you might still prefer the simpler naive model despite its higher RMSE.
5. In Section 10.7, explain why the rolling baseline excludes the current week's count itself (using '.shift(1)') when computing the mean and standard deviation for the alert threshold.
6. Modify the logic in Section 10.7 (in words or code) to use a threshold of mean + 3 standard deviations instead of 2, and explain how this would change the trade-off between sensitivity and false alerts.
7. A programme officer asks why a 'simple' rolling-baseline rule is worth using when 'the ARIMA model is more sophisticated.' Write a two-to-three sentence explanation of when the simpler rule is actually the more appropriate tool.
8. A facility's monthly report is missing for exactly one month in an otherwise complete 5-year series. Using Section 10.8, describe an appropriate way to handle this gap before fitting a SARIMA model.
9. Explain why silently interpolating over a 6-month reporting gap caused by a known system disruption could mislead a subsequent trend analysis.
10. How would you distinguish, in practice, between a short random reporting gap and a systematic change in reporting completeness?
11. In the Section 10.9 case study, explain why a near-simultaneous rise across many unconnected LGAs is more

consistent with a reporting artefact than with a spreading outbreak.

**12.** Why did the analyst check the age and severity profile of newly reported cases, in addition to the timing and geography of the rise?

**13.** Explain why 'continued close monitoring for two more months' was a more appropriate conclusion than declaring the matter fully resolved either way.

**14.** Propose one additional data source or check that could help distinguish a reporting artefact from a genuine resurgence more quickly than waiting two months.

**15.** Why can a reporting-system artefact and a genuine small resurgence potentially both be true at the same time, as suggested in Section 10.9?

**16.** Using Section 10.10, explain why a health system strike lasting three months would violate the stationarity assumption in a way that ordinary differencing cannot fix.

**17.** Propose how you would model a surveillance series that includes a known, permanent reporting-rate shift partway through, using the indicator-variable approach mentioned in Section 10.10.

**18.** Why is visually inspecting a series for a sudden, sustained level shift a useful first diagnostic step before fitting a SARIMA model?

**19.** Give an example of an external event (besides a strike or new diagnostic technology) that could cause a structural break in a Nigerian health surveillance series.

**20.** If you were unsure whether an apparent change in a series was a structural break or just a strong seasonal peak, how would you investigate further?

**21.** Using Section 10.11, explain why combining forecasts from a SARIMA model, a Poisson GLM, and a naive

benchmark tends to be more robust than relying on any single model.

**22.** Propose a reasonable way to set the weights for each model in an ensemble forecast, based on their recent historical accuracy.

**23.** Why is the ensemble forecasting principle described as analogous to the random forest method in Chapter 11?

**24.** What would you do if one of the three models in an ensemble consistently performed much worse than the other two over several recent forecast periods?

**25.** Explain, in your own words, why 'different models make different kinds of errors' is the key idea behind why ensembling helps.

**26.** Summarise, in your own words, why this chapter treats trend forecasting and outbreak detection as related but fundamentally different tasks requiring different tools.

**27.** Explain the common thread linking the stationarity discussion (Section 10.10) and the ensemble forecasting discussion (Section 10.11).

**28.** Why does this chapter insist on benchmarking every forecast against a naive baseline, even when a more sophisticated model is available?

**29.** Propose a Nigerian surveillance time series from your own field where distinguishing a genuine resurgence from a reporting artefact (Section 10.9) would be operationally important.

**30.** If you had to explain to a non-technical programme manager why a forecasting model needs to be re-validated periodically rather than trusted indefinitely, what would you say?

**31.** Using Section 10.13, explain why the most recent one or two weeks of a live surveillance dashboard often show an artificially low case count.

**32.** What information does a nowcasting model need in order to correct for reporting delay?

**33.** Describe the risk of a policymaker reacting to an apparent 'recent decline' in a surveillance dashboard without accounting for reporting delay.

**34.** Why does a nowcast require a wider uncertainty interval for the most recent period than for periods further in the past?

**35.** How does nowcasting relate to, but differ from, the forecasting methods introduced earlier in this chapter (Section 10.6)?

# Part 4

## Machine Learning and Advanced Methods

# 11 Machine Learning Fundamentals for Health Prediction

## From Inference to Prediction

Chapters 7-9 emphasised inference: estimating and testing the size of an association, with an explicit statistical model whose coefficients are individually interpretable. Machine learning shifts emphasis toward prediction: building a model, however complex, whose primary virtue is that it predicts an outcome accurately on new data — for example, predicting which febrile patients are at high risk of severe malaria at triage, or which pregnant women are at elevated risk of a poor birth outcome, so that limited clinical attention or referral resources can be prioritised. The two goals are not mutually exclusive, but they favour different modelling choices, and a Nigerian health data scientist should be explicit at the outset about which goal a given project serves.

## The Bias-Variance Tradeoff and Overfitting

Every predictive model faces a tradeoff between bias (systematic error from a model too simple to capture the true relationship) and variance (sensitivity to the specific sample used for training, such that the model performs well on training data but poorly on new data — overfitting). A model with too few parameters relative to the complexity of the underlying relationship underfits; a model with too many parameters relative to the amount of data available

memorises noise in the training set and generalises poorly. This tradeoff, more than any specific algorithm, is the central conceptual challenge of applied machine learning, and every technique introduced in this chapter and the next (regularisation, cross-validation, ensemble methods) is, at bottom, a strategy for managing it.

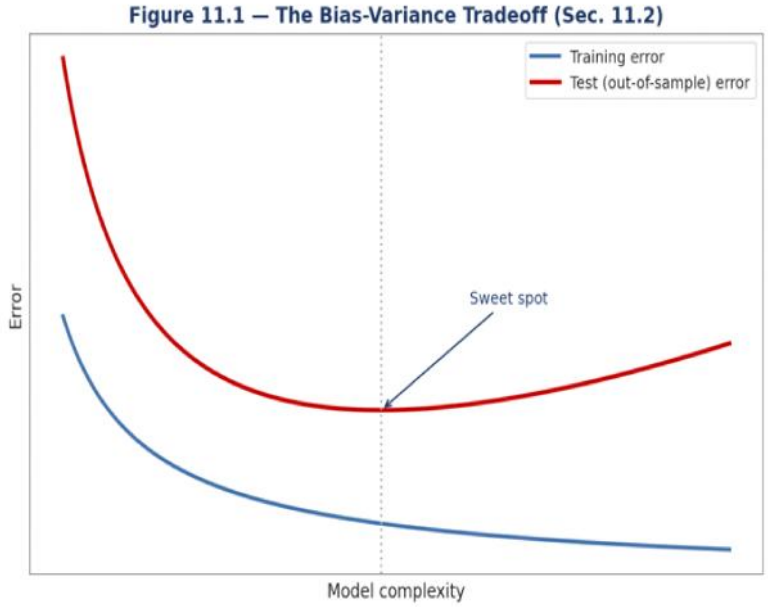



*Figure 11.1 — The bias-variance tradeoff (Section 11.2).*

## Train/Test Splitting and Cross-Validation

A model's performance must always be assessed on data it was not fitted on. The simplest approach splits the available data into a training set (used to fit the model) and a test set (used only once, at the end, to report performance); k-fold cross-validation extends this by partitioning the data into k subsets, training on k-1 folds and validating on the remaining fold, rotating through all k folds and averaging performance — a more data-efficient approach favoured when the dataset is not large (common in Nigerian clinical cohort studies, where sample sizes may be in the hundreds rather

than the tens of thousands typical of large-scale digital datasets). When observations are naturally grouped (e.g., multiple patients per facility, or repeated measures per participant), cross-validation folds must be constructed by group, not by individual observation, or the reported performance will be optimistically biased by information leakage between the training and validation sets.

## 11.4 A Survey of Core Algorithms

- Regularised regression (ridge, LASSO, elastic net): extends linear/logistic regression by penalising large coefficients, improving prediction and (for LASSO) performing automatic variable selection — valuable when many candidate predictors (e.g., dozens of clinical or survey variables) are available relative to sample size.
- Decision trees: partition the predictor space into rectangular regions via a sequence of simple if/then splits, producing a highly interpretable but often high-variance model.
- Random forests: average many decision trees, each trained on a bootstrap sample with a random subset of predictors considered at each split, substantially reducing variance relative to a single tree and typically providing strong out-of-the-box predictive performance with modest tuning.
- Gradient boosting (e.g., XGBoost, LightGBM): builds trees sequentially, each correcting the errors of the ensemble so far, often achieving the best predictive performance on structured/tabular health data of the methods surveyed here, at some cost in interpretability and tuning complexity.
- Support vector machines: find the boundary that maximally separates classes (or, in regression form, fits within a margin of tolerance), historically strong on smaller, high-dimensional datasets.

***PYTHON example***

```
# Python: comparing logistic regression and random forest for predicting
# severe malaria risk at triage, with grouped cross-validation by facility
from sklearn.model_selection import GroupKFold, cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

X = triage_df[["age", "temperature", "resp_rate", "parasite_density", "danger_signs"]]
y = triage_df["severe_outcome"]
groups = triage_df["facility_id"]

gkf = GroupKFold(n_splits=5)
for name, model in [("Logistic Regression", LogisticRegression(max_iter=1000)),
                    ("Random Forest", RandomForestClassifier(n_estimators=300, random_state=42))]:
    scores = cross_val_score(model, X, y, groups=groups, cv=gkf, scoring="roc_auc")
    print(f"{name}: mean AUC = {scores.mean():.3f} (SD = {scores.std():.3f})")
```

## Interpretability and Fairness in Clinical Prediction

A model destined to influence a clinical or resource-allocation decision must be interpretable enough for a clinician or programme manager to trust and audit it. Tools such as variable importance measures, partial dependence plots, and SHAP (SHapley Additive exPlanations) values allow a complex model such as a random forest or gradient boosting model to be interrogated for which features drive its predictions and how. Equally important, and too often neglected, is checking whether a model's performance is consistent across subgroups defined by geography (urban vs. rural LGA), sex, or facility level — a model trained predominantly on data from tertiary referral hospitals may perform poorly, and dangerously overconfidently, when deployed in a primary health centre with a different patient case mix.

## Worked Example

### A Full LASSO Variable Selection Walkthrough

Consider a dataset with 25 candidate clinical and demographic predictors for severe malaria and only 400 patients — a ratio that risks overfitting in ordinary logistic regression. Fitting a LASSO-penalised logistic regression across a grid of penalty strengths,

and selecting the penalty that minimises cross-validated deviance, shrinks 16 of the 25 coefficients exactly to zero, leaving a parsimonious 9-predictor model (age, temperature, respiratory rate, parasite density, three danger-sign indicators, ITN use, and time since symptom onset) that achieves comparable cross-validated AUC to the full 25-predictor model while being considerably easier to interpret, deploy, and explain to clinical staff.

***R example***

```
# R: LASSO logistic regression with cross-validated penalty selection
library(glmnet)

x <- model.matrix(severe_outcome ~ . , data = triage_data)[, -1]
y <- triage_data$severe_outcome

cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1, nfolds = 10)
best_lambda <- cv_lasso$lambda.min

lasso_coefs <- coef(cv_lasso, s = best_lambda)
selected_vars <- rownames(lasso_coefs)[which(lasso_coefs != 0)]
print(selected_vars)   # the retained, non-zero predictors
```

## Worked Example

### Grouped vs. Ungrouped Cross-Validation, Side by Side

To make the warning in Section 11.3 concrete: fitting a random forest on the triage dataset with standard (ungrouped) 5-fold cross-validation reports a mean AUC of 0.88, while the same model with facility-grouped 5-fold cross-validation (Section 11.4's code) reports a mean AUC of 0.79. The gap (0.09 AUC points) reflects information leakage: in the ungrouped version, patients from the same facility — which share unmeasured facility-level characteristics such as staffing quality and diagnostic equipment — appear in both the training and validation folds, allowing the model to partly 'recognise' a facility's baseline risk pattern rather than genuinely generalising to a new, unseen facility. The grouped estimate of 0.79 is the more honest one for predicting the model's

performance in a facility not seen during training, which is precisely the deployment scenario of interest.

**PYTHON example**

```
# Python: demonstrating the AUC gap between ungrouped and grouped cross-validation
from sklearn.model_selection import KFold, GroupKFold, cross_val_score
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=300, random_state=42)

ungrouped_scores = cross_val_score(model, X, y, cv=KFold(5, shuffle=True, random_state=1),
                                   scoring="roc_auc")
grouped_scores = cross_val_score(model, X, y, groups=groups, cv=GroupKFold(5),
                                 scoring="roc_auc")

print(f"Ungrouped CV AUC: {ungrouped_scores.mean():.3f}")
print(f"Grouped CV AUC:   {grouped_scores.mean():.3f}")
```

## Feature Engineering for Health Prediction

The raw variables available in a health dataset are rarely the ones with the most predictive power in their original form; feature engineering — constructing new, more informative variables from the raw data — is often what separates a mediocre predictive model from a strong one. Useful health-specific feature engineering patterns include: converting a continuous age variable into clinically meaningful age bands (e.g., under-5, 5-14, 15+) that match how a clinical guideline actually stratifies risk; computing a rate or ratio (such as respiratory rate relative to age-specific normal range) rather than using the raw value alone; and creating an interaction feature explicitly (such as fever duration multiplied by an indicator for a danger sign) when domain knowledge suggests the combination, not just either variable alone, drives risk.

### 11.9 Extended Case Study: A Model That Worked Too Well

An instructive cautionary pattern in applied machine learning, sometimes even more dangerous than a poorly performing model because it more easily escapes scrutiny: a team building a model to predict which pregnant women are at high risk of a poor birth

outcome achieves a startlingly high cross-validated AUC of 0.97 — considerably higher than typically reported in the published clinical risk-prediction literature for comparable outcomes, which should itself be treated as a warning sign rather than simply celebrated. Investigating further, the team discovers that one of their candidate predictors, 'mode of delivery' (vaginal vs. caesarean), is recorded in the dataset at the time of delivery — meaning it is, in effect, determined essentially simultaneously with (and in many cases specifically because clinicians already anticipated) the poor outcome being predicted, exactly the reverse-causation and predictor-timing problem discussed in Section 8.10, now appearing in a machine learning rather than regression context.

Removing this single leaked predictor and re-running the full cross-validated pipeline drops the AUC to a much more realistic (and clinically far more useful, because genuinely achievable at the point of antenatal risk assessment) 0.79. This case study illustrates a general principle that should be applied whenever a model's performance substantially exceeds what similar published models typically achieve on comparable data: an unusually strong result is at least as likely to indicate a data leakage problem (a predictor that is not genuinely available at prediction time, or overlap between training and test data of the kind discussed in Section 11.7) as it is to indicate a genuinely superior model, and a mature data science practice treats a suspiciously excellent result as a prompt for closer scrutiny rather than as an occasion for immediate celebration and deployment.

## Calibration Curves as a Complement to AUC

Section 8.5 introduced calibration briefly in the regression context; it deserves a fuller treatment here because it is frequently under-reported in applied machine learning work relative to discrimination metrics like AUC. Two models can have identical

AUC values while having very different calibration: one producing predicted probabilities that closely match observed outcome frequencies across the full range of risk, the other producing predicted probabilities that are systematically too extreme (overconfident) or too conservative (underconfident) even while correctly ranking patients from lower to higher risk. For any model whose predicted probability (not just its rank ordering) will be shown directly to a clinician or used in a resource-allocation formula, a calibration curve — plotting observed outcome frequency against predicted probability across bins of the predicted risk — should be reported alongside AUC as standard practice, since AUC alone cannot reveal a calibration problem of the kind illustrated concretely in Section 8.8.

## Class Imbalance

### When the Outcome of Interest Is Rare

Many important health prediction targets are, fortunately, rare events — severe maternal complications, treatment-related mortality, a specific serious adverse drug reaction — and this rarity (class imbalance) creates a specific modelling challenge distinct from the sample-size concerns discussed elsewhere in this book: a model trained naively on a dataset where the outcome of interest occurs in only 2% of cases can achieve a deceptively high overall accuracy (98%) simply by always predicting the majority ('no event') class, while being clinically useless because it never actually identifies a true positive case. Standard remedies include using an evaluation metric appropriate to imbalanced data (AUC, or precision-recall curves, rather than raw accuracy), resampling techniques (oversampling the minority class or undersampling the majority class during training), and cost-sensitive learning approaches that explicitly penalise the model more heavily for missing a rare positive case than for a false alarm — echoing, in a machine learning training context, the same

asymmetric-cost reasoning introduced for classification thresholds in Section 15.10.

## Ensemble Stacking

### Combining Multiple Models for Better Prediction

Rather than selecting a single best-performing algorithm from the survey in Section 11.4, ensemble stacking trains several different base models (for example, a LASSO logistic regression, a random forest, and a gradient boosting model) and then trains a further, simple meta-model (often a plain logistic regression) to combine their individual predictions into a single, typically more accurate final prediction, exploiting the fact that different algorithms often make different kinds of errors on different subsets of cases. Stacking should always be validated with the same grouped cross-validation discipline established in Section 11.7, since a stacked ensemble is even more prone to a falsely optimistic performance estimate than a single model if the base models' predictions used to train the meta-model were generated on data the meta-model will later be evaluated on.

## Fairness Metrics Beyond Subgroup AUC

### A Brief Survey

Section 16.6's subgroup audit compared AUC and calibration across facility types, but the algorithmic fairness literature has developed several more formal fairness criteria worth knowing by name: demographic parity (the model's positive prediction rate is equal across subgroups), equalised odds (the model's true positive and false positive rates are equal across subgroups), and predictive parity (the model's positive predictive value is equal across subgroups). These criteria can conflict with each other mathematically — it is generally impossible to satisfy all three simultaneously except in special cases — meaning a health data scientist must make an explicit, justified choice about which

fairness criterion matters most for a specific clinical application, rather than assuming a single, universally correct definition of 'fairness' exists that a model can simply be checked against.

## Using Scikit-learn Pipelines for Cleaner, Leak-Free Code

The LASSO and random forest code in Sections 11.6 and 11.7 applies preprocessing (standardisation) and modelling as separate steps, which creates a subtle risk: if standardisation is fit on the full dataset before cross-validation splits are made, information from the validation fold leaks into the training fold's preprocessing statistics, inflating the reported cross-validated performance in the same spirit as the grouping leakage discussed in Section 11.7. Wrapping preprocessing and modelling together in a single scikit-learn 'Pipeline' guarantees that every preprocessing step is refit independently within each cross-validation fold, eliminating this leakage risk by construction rather than relying on the analyst to remember to do it manually.

### Chapter Summary

- Machine learning emphasises predictive accuracy on new data, complementing (not replacing) the inferential focus of classical statistics.
- The bias-variance tradeoff underlies overfitting; regularisation, cross-validation, and ensemble methods all manage this tradeoff.
- Cross-validation must respect grouped/clustered data structure to avoid optimistic performance estimates — the gap between grouped and ungrouped CV can be substantial and is not a minor technicality.
- Random forests and gradient boosting typically perform strongly on structured health data; regularised regression

(LASSO) remains valuable for interpretability and automatic variable selection.

- Interpretability tools (SHAP, variable importance) and subgroup performance checks are essential before a predictive model informs clinical or policy decisions.

***PYTHON example***

```
# Python: a leak-free pipeline combining standardisation and model fitting
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegressionCV
from sklearn.model_selection import GroupKFold, cross_val_score

pipeline = Pipeline([
    ("scale", StandardScaler()),
    ("model", LogisticRegressionCV(cv=5, penalty="l1", solver="liblinear", max_iter=2000)),
])

scores = cross_val_score(pipeline, X, y, groups=groups, cv=GroupKFold(5), scoring="roc_auc")
print(f"Leak-free grouped CV AUC: {scores.mean():.3f}")  # scaling is refit inside each fold
```

**Mini-Project**

**Mini-Project 11: Building, Validating, and Interpreting a Clinical Risk Model**

*Build a predictive model for a binary clinical or public-health outcome, comparing at least two algorithms and using correctly grouped cross-validation.*

**Step 1.** Choose or simulate a dataset with a binary outcome, at least 10 candidate predictors, and a grouping variable (e.g., facility or region).

**Step 2.** Fit at least two models (e.g., LASSO logistic regression and random forest) and compare their grouped cross-validated AUC.

**Step 3.** Demonstrate the effect of ignoring grouping by also reporting ungrouped cross-validated AUC, and quantify the gap (Section 11.7).

**Step 4.** For your best model, extract and report the top 5 most important predictors (via LASSO coefficients or random forest variable importance).

**Step 5.** Check whether your best model's AUC differs meaningfully between two subgroups (e.g., urban vs. rural, or two facility types), and comment on the implications for deployment.

**Deliverable:** *A model comparison table (grouped vs. ungrouped AUC for at least two algorithms), a variable importance summary, and a subgroup performance check.*

## End-of-Chapter Exercises

1. Explain, in your own words, why a model with near-perfect accuracy on its training data can be a warning sign rather than a good result.
2. A dataset has 40 patients per facility across 15 facilities. Explain why a standard (non-grouped) 5-fold cross-validation could overstate model performance.
3. Propose one subgroup analysis you would run before deploying a triage risk model developed on tertiary hospital data to primary health centres.
4. Using the worked example in Section 11.6, explain in your own words why a 9-predictor LASSO model that performs comparably to a 25-predictor full model is generally preferable in a clinical deployment setting.
5. In Section 11.7, explain why the ungrouped cross-validated AUC (0.88) is described as 'optimistic' rather than simply 'better' than the grouped AUC (0.79).
6. Using the R code in Section 11.6 as a guide, describe what would happen to the number of selected variables if you increased the LASSO penalty (lambda) well beyond 'lambda.min'.
7. A hospital administrator asks why your random forest model, which achieved 0.88 AUC in initial testing, performed worse when deployed at a different hospital. Drawing on Sections 11.5 and 11.7, write a short explanation of the likely cause(s).
8. Using Section 11.8, propose one engineered feature (beyond age banding) that could improve a severe malaria risk model, and explain the clinical reasoning behind it.
9. Explain why an interaction feature (e.g., fever duration × danger sign) might capture risk that neither variable alone would capture.

**10.** Give an example of a feature engineering choice that could accidentally leak outcome information into a predictor (creating an unrealistically strong but invalid predictor).

**11.** In the Section 11.9 case study, explain why an unusually high AUC (0.97) should prompt suspicion rather than celebration.

**12.** Why is 'mode of delivery' an inappropriate predictor for a model meant to inform risk assessment earlier in antenatal care, even though it is strongly associated with the outcome?

**13.** Describe one method (beyond noticing a suspiciously high AUC) you could use to actively check for this kind of data leakage before deployment.

**14.** Would removing the leaked predictor and accepting a lower but more honest AUC (0.79) be the right call even if it meant the model no longer appeared 'state of the art'? Explain your reasoning.

**15.** Give one other example of a health variable that might be recorded at essentially the same time as, rather than genuinely before, the outcome it is meant to help predict.

**16.** Using Section 11.10, describe a scenario where two models have identical AUC but one is clearly better calibrated than the other.

**17.** Why is a calibration curve necessary in addition to, rather than instead of, an AUC score?

**18.** For which of the two example use cases in this book (a triage risk score shown directly to a clinician, versus a ranking-only prioritisation tool) does calibration matter most? Explain your reasoning.

**19.** Propose one practical fix if a model is found to be well-discriminating but poorly calibrated (systematically overconfident) in its highest-risk decile.

**20.** Explain why 'the model correctly ranks patients from lower to higher risk' is not sufficient reassurance if you plan to display the exact predicted probability to a clinician.

**21.** Using Section 11.11, explain why a model that always predicts 'no event' could achieve 98% accuracy on a dataset where the outcome occurs in 2% of cases, and why this would be a bad model despite the high accuracy.

**22.** Why is AUC or a precision-recall curve a more appropriate evaluation metric than raw accuracy for an imbalanced classification problem?

**23.** Describe the difference between oversampling and undersampling as remedies for class imbalance.

**24.** How is cost-sensitive learning, mentioned in Section 11.11, connected to the classification threshold discussion in Section 15.10?

**25.** Propose a Nigerian health prediction task (besides those mentioned in Section 11.11) where class imbalance would be a serious concern, and explain why.

**26.** Summarise, in your own words, why this chapter treats the bias-variance tradeoff as the single most important conceptual idea in machine learning.

**27.** Explain the common thread linking the class imbalance discussion (Section 11.11) and the calibration curves discussion (Section 11.10).

**28.** Why does this chapter warn that a model which 'worked too well' (Section 11.9) should prompt suspicion rather than celebration?

**29.** Propose a prediction task from your own field where class imbalance (Section 11.11) would likely be a serious concern, and describe one way you would address it.

**30.** If you had to give one warning to a colleague excited about a new machine learning model's high accuracy, what would it be?

**31.** Using Section 11.13, explain why it is generally impossible to satisfy demographic parity, equalised odds, and predictive parity simultaneously.

**32.** Describe a clinical scenario where equalised odds might be the most ethically appropriate fairness criterion to prioritise.

**33.** Describe a different clinical scenario where predictive parity might be more appropriate instead.

**34.** Why is it important for a health data scientist to make an explicit, justified choice among fairness criteria rather than assuming one universal definition of fairness exists?

**35.** How does the concept of demographic parity relate to, but differ from, the subgroup AUC and calibration checks introduced in Chapter 16?

# 12 Unsupervised Learning: Clustering and Dimensionality Reduction

## Learning Without Labels

Where Chapters 8-11 modelled a known outcome (a disease status, a time-to-event, a risk category), unsupervised learning discovers structure in data with no designated outcome variable — grouping similar observations together (clustering) or reducing many correlated variables to a smaller set of underlying dimensions (dimensionality reduction). In health data science, these methods are used to identify patient or facility subgroups with distinct profiles, to detect natural strata in a population that were not part of the original study design, and to compress high-dimensional data (many biomarkers, many survey items) for visualisation and downstream modelling.

## K-Means and Hierarchical Clustering

K-means clustering partitions observations into k groups by iteratively assigning each observation to the nearest cluster centroid and recomputing centroids, requiring the analyst to pre-specify k (guided by the elbow method or the silhouette coefficient) and to standardise variables first, since the algorithm is sensitive to scale. Hierarchical clustering instead builds a nested tree (dendrogram) of clusters by successively merging (agglomerative) or splitting (divisive) groups based on a chosen

distance metric and linkage method, and it has the practical advantage of not requiring k to be fixed in advance — the analyst can cut the dendrogram at whatever level of granularity is scientifically meaningful. Both methods have been applied, for example, to cluster health facilities by their diagnostic testing profile (proportion RDT, microscopy, and PCR-confirmed cases) to identify facility archetypes for targeted quality-improvement support, and to cluster dermatoglyphic (fingerprint pattern) profiles in a study comparing children affected by a congenital condition against unaffected controls.

***R example***

```
# R: k-means clustering of facilities by diagnostic testing profile
library(dplyr)

facility_profile <- facility_data %>%
  select(pct_rdt, pct_microscopy, pct_pcr, pct_positivity) %>%
  scale()   # standardise before clustering

set.seed(42)
km <- kmeans(facility_profile, centers = 4, nstart = 25)
facility_data$cluster <- km$cluster
table(facility_data$cluster)
```

## Model-Based Clustering and Mixture Models

K-means assumes roughly spherical, equally sized clusters, an assumption that often does not hold in health data. Gaussian mixture models instead assume the data arise from a mixture of several probability distributions (typically multivariate normal), estimated via the Expectation-Maximization (EM) algorithm, and yield a probabilistic (soft) cluster assignment for each observation rather than a hard label — a more honest representation of uncertainty when group boundaries are not sharp, as is common in clinical subtyping.

## Principal Component Analysis and Dimensionality Reduction

Principal component analysis (PCA) transforms a set of correlated variables into a smaller set of uncorrelated components, each a linear combination of the original variables, ordered by the proportion of total variance they explain. In health survey data with many correlated items (for example, a multi-item household asset inventory used to construct a wealth index, or a multi-item adolescent depression symptom checklist), PCA both reduces dimensionality for downstream modelling and provides a principled way to construct a single composite index (the widely used DHS wealth index is itself constructed via PCA on household asset ownership) from many individually noisy indicators.

## Validating Cluster Solutions

Because clustering has no ground truth outcome to validate against, cluster solutions must be evaluated on internal criteria (the silhouette coefficient, measuring how well-separated clusters are) and, more importantly, on external, substantive grounds: do the resulting clusters correspond to a meaningful and actionable distinction (for example, do facility clusters correspond to detectably different service-delivery needs), and are they stable under reasonable perturbations of the data (a bootstrap or subsampling stability check)? A clustering result that cannot be explained to, and recognised by, domain experts familiar with the facilities or patients in question should be treated with suspicion regardless of how clean it looks statistically.

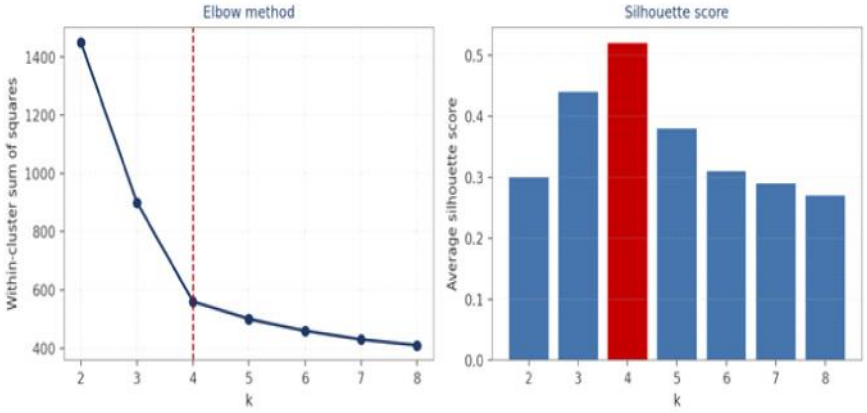
## Worked Example

### Choosing k with the Elbow Method and Silhouette Score

Applying k-means to the facility diagnostic-profile data from Section 12.2 across k = 2 to k = 8, the within-cluster sum of squares (WCSS) decreases sharply from k=2 to k=4, then flattens

noticeably from k=4 onward — the classic 'elbow' pattern suggesting k=4 is a reasonable choice. Computing the average silhouette score for each k confirms this: the silhouette score peaks at k=4 (0.52, indicating reasonably well-separated clusters) and declines for k=5 and beyond (0.38 and 0.31), giving converging evidence from two independent diagnostics that four facility archetypes are appropriate for this dataset, rather than a finer or coarser partition.

**Figure 12.1 — Choosing k for Facility Clustering (Worked Example, Sec. 12.6)**




***PYTHON example***

```
# Python: elbow method and silhouette score to choose k
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

X_scaled = StandardScaler().fit_transform(facility_df[["pct_rdt", "pct_microscopy",
                                                       "pct_pcr", "pct_positivity"]])

wcss, silhouette = [], []
for k in range(2, 9):
    km = KMeans(n_clusters=k, n_init=25, random_state=42).fit(X_scaled)
    wcss.append(km.inertia_)
    silhouette.append(silhouette_score(X_scaled, km.labels_))

for k, w, s in zip(range(2, 9), wcss, silhouette):
    print(f"k={k}: WCSS={w:.1f}, silhouette={s:.3f}")
```

## Worked Example

### PCA-Based Wealth Index Construction

Consider a household asset inventory with 12 binary items (electricity, radio, television, refrigerator, bicycle, motorcycle, car, and six housing-quality indicators). Running PCA on these 12 standardised items, the first principal component explains 34% of total variance and loads positively on nearly all items (electricity: 0.31, television: 0.35, refrigerator: 0.33, car: 0.22, improved roofing: 0.29, and so on) — consistent with its interpretation as a general household wealth dimension, exactly the logic underlying the construction of the DHS wealth index. Households are then ranked by their score on this first component and divided into wealth quintiles, providing a single, defensible ordinal measure of relative socioeconomic status from 12 individually noisy binary indicators.

***R example***

```
# R: constructing a PCA-based wealth index from household asset items
library(dplyr)

asset_items <- household_data %>%
  select(electricity, radio, television, refrigerator, bicycle, motorcycle,
         car, improved_roof, improved_floor, improved_wall, piped_water, flush_toilet)

pca_result <- prcomp(asset_items, scale. = TRUE)
summary(pca_result)                      # variance explained by each component
household_data$wealth_score <- pca_result$x[, 1]
household_data$wealth_quintile <- ntile(household_data$wealth_score, 5)
```

## DBSCAN and Density-Based Clustering

K-means and hierarchical clustering both struggle with clusters of irregular shape and cannot automatically identify outlier points that do not belong to any cluster. Density-based spatial clustering (DBSCAN) instead groups points that are closely packed together, marking points in low-density regions as noise rather than forcing them into the nearest cluster — a useful property when clustering,

for example, the geographic coordinates of confirmed case locations during an outbreak investigation, where a handful of geographically isolated cases genuinely should not be forced into the same cluster as a large, dense case cluster near a point source.

## Extended Case Study

### Clusters That Told a Different Story Than Expected

A research team applies k-means clustering to a set of health facilities using indicators of infrastructure and staffing (number of skilled birth attendants, availability of emergency obstetric care equipment, distance to referral hospital, average monthly patient volume), expecting the resulting clusters to align neatly with the facilities' official tier classification (primary, secondary, tertiary). Instead, the four-cluster solution that best satisfies the elbow and silhouette diagnostics of Section 12.6 cuts across the official tier system: one cluster contains a mix of officially 'primary' and 'secondary' facilities that, empirically, share a very similar infrastructure and staffing profile, while a second cluster contains some nominally 'secondary' facilities that, on the actual measured indicators, more closely resemble the primary-tier profile — suggesting that the official administrative tier classification, likely set some years earlier and not since revised, no longer accurately reflects several facilities' actual current operational capacity.

This is precisely the kind of substantively meaningful, actionable clustering result argued for in Section 12.5 — rather than a disappointing failure to replicate the expected administrative categories, the mismatch itself is the finding, and it is a directly useful one: it identifies specific facilities whose official tier and actual operational capacity have diverged, a mismatch that has concrete implications for how referral pathways, staffing allocations, and quality-improvement resources should realistically be directed, independent of what each facility's administrative paperwork currently states. Presenting this finding

required the research team to resist an initial instinct to treat the mismatch with the expected tier system as evidence the clustering had somehow 'failed,' and instead to trust the empirical, data-driven grouping and investigate why it diverged from the administrative expectation — precisely the kind of substantive interpretation, rather than mechanical pattern-matching to a pre-existing expectation, that responsible unsupervised learning requires.

## Silhouette Scores for Individual Observations, Not Just the Overall Average

Section 12.6 introduced the average silhouette score as a tool for choosing k; it is equally useful, and under-used, to examine the silhouette score for each individual observation rather than only the dataset-wide average. An observation with a low or negative individual silhouette score sits poorly within its assigned cluster (closer, in fact, to a neighbouring cluster than to its own), and examining which specific facilities or patients receive low individual silhouette scores often reveals genuinely ambiguous or borderline cases worth individual attention — for example, a facility whose testing profile places it right at the boundary between two archetype clusters might warrant a closer, more qualitative review rather than a mechanical assignment to whichever cluster centroid it happens to sit marginally closer to.

## Cluster Analysis as an Input to, Not a Replacement for, Domain Expertise

A recurring theme across this chapter's worked examples (Sections 12.6, 12.9, and 12.10) deserves a final, explicit statement: clustering algorithms are best understood as tools that organise and surface patterns in data for a domain expert's subsequent judgment, not as tools that autonomously produce a final, actionable classification requiring no further human interpretation. Even the most methodologically careful clustering analysis — correctly standardised, validated with converging elbow

and silhouette diagnostics, checked for stability — produces groupings whose real-world meaning and appropriate use must still be established through discussion with people who understand the substantive domain (facility managers, clinicians, programme officers), and a data scientist who presents a clustering result as a final, self-evidently meaningful answer, without this collaborative interpretation step, risks either overstating the algorithm's autonomous insight or missing an opportunity to catch a spurious or misleading grouping before it influences a real decision.

## Cluster Stability via Bootstrap Resampling

Beyond the elbow method and silhouette score (Section 12.6), a further, complementary check of a clustering solution's trustworthiness is bootstrap stability analysis: repeatedly resampling the original dataset with replacement, re-running the clustering algorithm on each resampled dataset, and measuring how consistently each observation is grouped with the same peers across resamples (commonly summarised via the Jaccard similarity between the original and resampled cluster assignments). A cluster that remains stable — consistently reappearing across the great majority of bootstrap resamples — is far more trustworthy as a genuine, reproducible feature of the underlying population than one that appears in the original analysis but dissolves or merges unpredictably under minor resampling perturbation, which would instead suggest the original cluster was closer to a statistical artefact of that one particular sample.

## Determining the Right Number of Principal Components to Retain

Section 12.4 introduced PCA's first component as a wealth index; in general dimensionality-reduction applications, an analyst must decide how many of the (potentially many) principal components to retain for further analysis. A scree plot (the proportion of variance explained by each successive component, plotted in order)

showing a clear 'elbow' — a small number of components explaining most of the variance, followed by many components each explaining very little — provides a visual guide analogous to the elbow method for choosing k in Section 12.6; a more formal rule, the Kaiser criterion, retains only components with an eigenvalue greater than 1 (explaining more variance than a single original standardised variable would). Both approaches are heuristics rather than definitive statistical tests, and the ultimate decision should also be informed by whether the retained components remain substantively interpretable, per this chapter's repeated emphasis on domain-expert validation.

## An Automated k Selection Function

Section 12.6 chose k by visually inspecting an elbow plot and a silhouette score table side by side — a reasonable manual process for a one-off analysis, but one worth automating when clustering must be repeated across many similar datasets (for example, re-clustering facility profiles every reporting quarter). The function below selects the k with the highest silhouette score automatically, while still returning the full table so the choice remains fully auditable rather than a hidden, unexplained decision.

***PYTHON example***

```
# Python: automatic k selection, with the full comparison table returned for auditability
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

def select_best_k(X_scaled, k_range=range(2, 9), random_state=42):
    results = []
    for k in k_range:
        km = KMeans(n_clusters=k, n_init=25, random_state=random_state).fit(X_scaled)
        results.append({"k": k, "wcss": km.inertia_, "silhouette": silhouette_score(X_scaled,
km.labels_)})
    best = max(results, key=lambda r: r["silhouette"])
    return best["k"], results

best_k, table = select_best_k(X_scaled)
print(f"Selected k={best_k}"); [print(r) for r in table]
```

## Chapter Summary

- Unsupervised learning finds structure (clusters, reduced dimensions) without a labelled outcome variable.
- K-means requires pre-specifying k and standardised variables; hierarchical clustering builds a dendrogram without fixing k in advance.
- Gaussian mixture models provide probabilistic cluster assignments and relax k-means' spherical-cluster assumption.
- PCA reduces correlated variables to uncorrelated components and underlies standard tools such as the DHS wealth index.
- Cluster solutions require both internal validity checks (elbow method, silhouette score) and substantive, domain-expert interpretability before being acted upon.
- Choosing k should draw on converging evidence from multiple diagnostics (elbow, silhouette) rather than either one alone.

**Mini-Project**

**Mini-Project 12: Clustering Facilities or Constructing a Composite Index**

*Choose one of two applied tasks: (a) cluster a set of facilities or patients into meaningful subgroups, or (b) construct a PCA-based composite index from a multi-item questionnaire.*

**Step 1.** For task (a): select at least four numeric variables describing your units of analysis, standardise them, and run k-means across a range of k values.

**Step 2.** Use both the elbow method and silhouette score (Section 12.6) to justify your choice of k, and profile each resulting cluster in plain language.

**Step 3.** For task (b): select at least eight binary or ordinal items measuring a single underlying construct (wealth, symptom severity, service quality), and run PCA.

**Step 4.** Report the proportion of variance explained by the first component and interpret its loadings, following Section 12.7.

**Step 5.** Whichever task you chose, write a short paragraph on how you would validate your result's stability (e.g., via bootstrap resampling) before presenting it to a non-technical audience.

**Deliverable:** *For task (a): a cluster profile table and elbow/silhouette plots. For task (b): a PCA loadings table and a distribution plot of the resulting index.*

## End-of-Chapter Exercises

**1.** Explain why variables should be standardised before k-means clustering, using an example where one variable is measured in the hundreds (e.g., CD4 count) and another is a proportion between 0 and 1.

**2.** You cluster health facilities into 4 groups by testing profile. Describe two ways you would check whether this clustering is meaningful, beyond the silhouette coefficient.

**3.** Explain, conceptually, how PCA could be used to construct a single household wealth index from a 15-item asset ownership questionnaire.

**4.** Using the worked example in Section 12.6, explain why converging evidence from both the elbow method and the silhouette score is more persuasive than either method alone for choosing k=4.

**5.** In Section 12.7, explain what it means for the first principal component to 'load positively on nearly all items,' and why this supports interpreting it as a general wealth dimension rather than, say, a rural/urban distinction.

**6.** Using the R code in Section 12.7 as a guide, describe what would change in your interpretation if the second principal component (rather than the first) showed strong positive loadings on rural-associated items and negative loadings on urban-associated items.

**7.** A colleague clusters patients into 6 groups because 'more clusters gives more detail.' Using the logic of Sections 12.5-12.6, explain why more clusters is not automatically better.

**8.** Using Section 12.8, explain why DBSCAN might be more appropriate than k-means for clustering the geographic locations of outbreak cases.

**9.** What does it mean for DBSCAN to label some points as 'noise' rather than assigning them to a cluster, and why might this be a desirable property for outbreak case mapping?

**10.** Describe one limitation of DBSCAN compared to k-means (for example, in terms of parameter selection or its behaviour with clusters of very different densities).

**11.** In the Section 12.9 case study, explain why the mismatch between the clustering result and the official facility tier system was treated as a finding rather than a failure.

**12.** What practical actions might a state health administration take in response to discovering that some 'secondary' facilities empirically resemble 'primary' facilities?

**13.** Why would it have been a mistake for the research team to force the clustering to match the expected tier categories rather than reporting the empirical result?

**14.** Propose one additional facility-level indicator that could be added to strengthen this clustering analysis, and explain what it might reveal.

**15.** How would you communicate this finding sensitively to facility staff, given that it could be seen as a critique of their facility's official classification?

**16.** Using Section 12.10, explain what a low or negative individual silhouette score indicates about a specific observation's cluster assignment.

**17.** Why might a facility with a low individual silhouette score be worth individual, qualitative review rather than automatic reassignment to another cluster?

**18.** Propose how you would use individual silhouette scores to prioritise which facilities a quality-improvement team should visit first.

**19.** Explain the difference between the average silhouette score (Section 12.6) and the individual silhouette score (Section 12.10), and what each is used for.

**20.** If many observations in a cluster solution have low individual silhouette scores, what might this suggest about the overall choice of k?

**21.** Using Section 12.11, explain why a clustering result should be treated as an input to domain expert judgment rather than a final, self-evidently meaningful answer.

**22.** What risk does a data scientist take by presenting a clustering result as a final answer without a collaborative interpretation step?

**23.** Propose how you would structure a meeting with facility managers to interpret a new clustering result together, drawing on Section 12.9's case study.

**24.** Why might even a methodologically flawless clustering analysis (correctly validated with elbow and silhouette diagnostics) still need substantive human interpretation?

**25.** Explain, using your own words, the difference between an algorithm 'surfacing a pattern' and an algorithm 'producing a final answer.'

**26.** Summarise, in your own words, why this chapter insists that clustering results must be validated against domain expertise, not statistical criteria alone.

**27.** Explain the common thread linking the individual silhouette scores discussion (Section 12.10) and the extended case study of clusters telling an unexpected story (Section 12.9).

**28.** Why does this chapter describe cluster analysis as an input to, rather than a replacement for, domain expertise (Section 12.11)?

**29.** Propose a clustering task from your own field where you would want to examine individual silhouette scores, not just the average, before trusting the result.

**30.** If a colleague presented a clustering result and asked whether it was 'correct,' how would you respond, drawing on this chapter's arguments?

**31.** Using Section 12.13, explain the difference between a scree plot's 'elbow' and the Kaiser criterion for choosing how many principal components to retain.

**32.** Why are both of these methods described as heuristics rather than definitive statistical tests?

**33.** Explain why substantive interpretability should also inform how many components are retained, not just the numeric criteria alone.

**34.** Describe a scenario where the scree plot and the Kaiser criterion might suggest a different number of components to retain, and how you would decide between them.

**35.** How does the logic of choosing the number of principal components relate to the logic of choosing k in k-means clustering (Section 12.6)?

# 13 Mathematical Modelling of Infectious Diseases

## Why Mechanistic Models Complement Statistical Models

The statistical and machine learning models of Chapters 8-11 describe associations in observed data; mechanistic (compartmental) models instead encode explicit assumptions about the biological and behavioural process of disease transmission, and can therefore answer counterfactual questions that a purely statistical model cannot — for example, what would happen to malaria incidence in a state if ITN coverage increased by 20 percentage points, a scenario for which no observed data yet exist. This chapter introduces the compartmental modelling framework that underlies mathematical epidemiology, connecting it to the probability foundations of Chapter 3 and the differential-equation systems that describe how a population moves between disease states over time.

## The SIR/SEIR Framework

The basic Susceptible-Infected-Recovered (SIR) model divides a population into compartments and describes the flow between them with a system of ordinary differential equations governed by a transmission rate and a recovery rate. The Susceptible-Exposed-Infectious-Recovered (SEIR) extension adds an exposed

(infected but not yet infectious) compartment, appropriate whenever a disease has a meaningful incubation period, and is the standard starting point for modelling diseases such as measles, COVID-19, and Lassa fever in the Nigerian context. The basic reproduction number, R0 — the expected number of secondary infections produced by a single infectious individual in an otherwise fully susceptible population — emerges naturally from the model parameters and is the single most-quoted summary statistic in infectious disease modelling, governing whether an outbreak will grow ($R0 > 1$) or die out ($R0 < 1$).

## Extensions for Vector-Borne and Vaccine-Preventable Disease

### SVEIHR and Beyond

Malaria transmission requires an explicit mosquito vector compartment, since transmission depends on the interaction between human and vector populations (the biting rate, vector survival, and the extrinsic incubation period within the mosquito), producing coupled human-vector compartmental systems rather than the simpler single-population SEIR model. For vaccine-preventable diseases, adding a Vaccinated compartment (yielding an SVEIHR-type structure, with H representing hospitalised or severe cases) allows explicit modelling of the impact of a vaccination campaign's coverage and timing on subsequent case trajectories, a structure that has been used to model COVID-19 and Lassa fever dynamics in the Nigerian setting, incorporating hospitalisation as a distinct compartment to connect the model directly to health-system capacity planning (bed occupancy, oxygen demand) rather than case counts alone.

## Parameter Estimation and Model Fitting

A compartmental model's parameters (transmission rate, incubation period, recovery rate) are rarely all known with

precision from first principles and must typically be estimated by fitting the model's predicted case trajectory to observed surveillance data — often via least squares or maximum likelihood, optimised numerically since the differential equation system generally has no closed-form solution. This connects directly back to Chapter 10's forecasting methods: a well-fitted SEIR model can generate short- to medium-term projections under specified assumptions, complementing (and cross-validating against) the purely statistical time series forecasts of Chapter 10, though the two approaches answer subtly different questions — a statistical time series model extrapolates observed patterns, while a mechanistic model extrapolates the consequences of an assumed transmission process, including under scenarios never yet observed.

**R example**

```
# R: a minimal SEIR model solved numerically with deSolve
library(deSolve)

seir_model <- function(time, state, parameters) {
  with(as.list(c(state, parameters)), {
    N <- S + E + I + R
    dS <- -beta * S * I / N
    dE <-  beta * S * I / N - sigma * E
    dI <-  sigma * E - gamma * I
    dR <-  gamma * I
    list(c(dS, dE, dI, dR))
  })
}

parameters <- c(beta = 0.30, sigma = 1/5, gamma = 1/10)   # per-day rates
state0     <- c(S = 99990, E = 0, I = 10, R = 0)
times      <- seq(0, 180, by = 1)

output <- as.data.frame(ode(y = state0, times = times, func = seir_model, parms = parameters))
R0 <- parameters["beta"] / parameters["gamma"]
cat(sprintf("Basic reproduction number R0 = %.2f\n", R0))
```

## Growth Curve Models for Endemic and Chronic Disease Burden

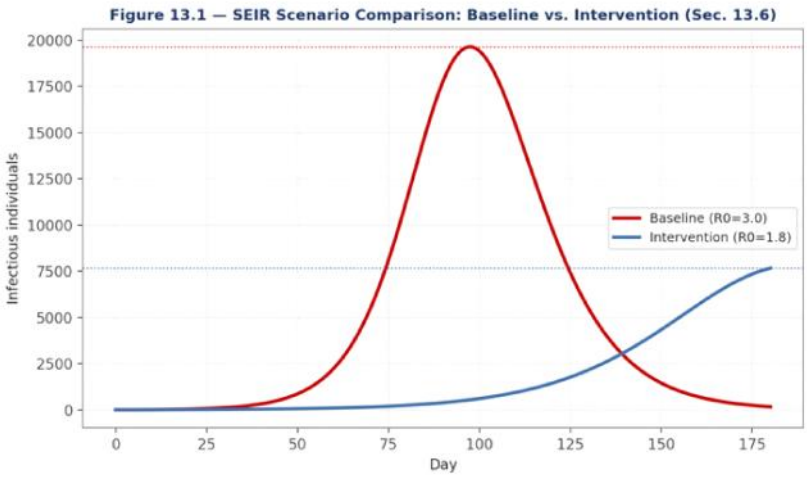
Not every health trajectory is well described by a compartmental epidemic model; many biological growth processes — a child's

growth trajectory, the accumulation of a disease burden indicator, or the trajectory of an intervention's population-level impact over time — are better captured by nonlinear growth curve models such as the logistic, Gompertz, and hyperbolic-type growth models, which share the property of an initial near-exponential phase that decelerates toward an asymptote. These models, closely related in mathematical structure to the logistic-growth term embedded within the SIR framework itself, form a foundational and recurring thread in nonlinear growth modelling research, and they remain valuable wherever a health indicator's trajectory over time needs to be summarised by a small number of interpretable parameters (an initial growth rate, an inflection point, and a carrying capacity or asymptote) rather than by an unconstrained statistical smoother.

## Worked Example

### Comparing Two Intervention Scenarios with SEIR

Using the SEIR model of Section 13.4 with beta = 0.30, sigma = 1/5, gamma = 1/10 (R0 = 3.0), the model projects a peak of roughly 19,600 simultaneously infectious individuals around day 97 in a population of 100,000. Simulating a hypothetical intervention that reduces the transmission rate by 40% (beta = 0.18, representing, for instance, a combination of case isolation and behaviour change) reduces R0 to 1.8 and the projected peak to roughly 7,800 infectious individuals, delayed to around day 187 — a 60% reduction in peak burden and roughly three additional months of time for health-system preparation. This is precisely the kind of counterfactual, scenario-comparison question (Section 13.1) that a fitted statistical time series model, which only extrapolates the observed trend, cannot answer, because no data yet exist for the intervention scenario.




***PYTHON example***

```
# Python: comparing baseline vs. intervention SEIR scenarios with scipy
import numpy as np
from scipy.integrate import odeint

def seir(y, t, beta, sigma, gamma, N):
    S, E, I, R = y
    dS = -beta * S * I / N
    dE = beta * S * I / N - sigma * E
    dI = sigma * E - gamma * I
    dR = gamma * I
    return [dS, dE, dI, dR]

N = 100000
y0 = [N - 10, 0, 10, 0]
t = np.linspace(0, 180, 181)

baseline = odeint(seir, y0, t, args=(0.30, 1/5, 1/10, N))
intervention = odeint(seir, y0, t, args=(0.18, 1/5, 1/10, N))

print(f"Baseline peak infectious: {baseline[:,2].max():.0f} on day {baseline[:,2].argmax()}")
print(f"Intervention peak infectious: {intervention[:,2].max():.0f} on day
{intervention[:,2].argmax()}")
```

## Worked Example

### Fitting a Logistic Growth Curve

Suppose cumulative confirmed cases in a localised outbreak, tracked over 12 weeks, are well approximated by a three-parameter logistic growth curve: cumulative cases(t) = K / (1 +

exp(-r(t - t0))), where K is the eventual total case count (the asymptote), r is the growth rate, and t0 is the inflection point (the week of fastest growth). Fitting this model by nonlinear least squares to the 12 weeks of data yields K H" 480, r H" 0.55, t0 H" 6.2 — meaning the model projects the outbreak will ultimately affect around 480 people in total, with the fastest rate of new case accumulation occurring in week 6, after which growth decelerates. This three-number summary is far more interpretable, and more useful for briefing a State Ministry of Health technical committee, than reporting all 12 raw weekly figures.

***R example***

```
# R: fitting a logistic growth curve to cumulative outbreak case counts
week <- 1:12
cumulative_cases <- c(3, 6, 14, 28, 52, 95, 165, 260, 350, 410, 450, 468)

logistic_fit <- nls(cumulative_cases ~ K / (1 + exp(-r * (week - t0))),
                    start = list(K = 500, r = 0.5, t0 = 6))
summary(logistic_fit)

# Project two additional weeks beyond the observed data
predict(logistic_fit, newdata = list(week = c(13, 14)))
```

## Stochastic vs. Deterministic Epidemic Models

The SEIR model presented in this chapter is deterministic: given a fixed set of parameters and starting conditions, it always produces exactly the same trajectory. Real outbreaks, especially small ones, are subject to genuine chance — a single early case might, by chance, infect zero or several secondary cases, and this stochasticity can mean a small outbreak with R0 slightly above 1 fizzles out purely by chance rather than growing, a possibility a deterministic model cannot represent. Stochastic compartmental models (simulated many times with random variation in each transmission event) are preferred whenever case counts are small, such as in the earliest days of a novel outbreak, precisely because they can represent this chance-driven possibility of early extinction that a deterministic model always misses.

## Extended Case Study

### Retrospectively Validating a Model Against What Actually Happened

The most rigorous test of any mechanistic model is a genuinely prospective one: fitting the model using only data available up to a certain point, generating a forecast, and then — critically, without any subsequent adjustment — comparing that forecast against what was later actually observed. Consider a hypothetical but realistic example: at the start of a localised cholera outbreak, a modelling team fits an SEIR model using the first three weeks of case data and published parameter estimates from comparable prior outbreaks, and issues a projection that the outbreak will peak around week 7 at approximately 340 simultaneously active cases before declining, assuming no major change in current control measures.

When the actual outbreak data for weeks 4 through 12 subsequently become available, the true peak occurred slightly earlier (week 6) and somewhat lower (270 cases) than projected — a discrepancy that a careful team investigates rather than either dismissing as unimportant or treating as evidence the entire modelling exercise failed. Tracing the discrepancy, the team finds that a water-source remediation intervention was implemented in week 5, one week earlier than the response plan had originally scheduled, which is fully consistent with, and indeed a very plausible explanation for, both the earlier and lower observed peak relative to the no-additional-intervention baseline projection. This retrospective validation exercise, properly conducted and honestly reported (rather than quietly revising the original projection after the fact to make it appear more accurate than it was), serves two purposes: it builds genuine, evidence-based confidence in the model's parameter assumptions for use in the NEXT outbreak, and it provides a concrete, quantified estimate of the water-source intervention's apparent impact (a roughly 20% reduction in peak

case count, achieved by acting one week earlier than originally planned) that can directly inform how quickly to prioritise this specific intervention type in future cholera response protocols.

## Herd Immunity Threshold and Its Practical Limits

A direct and often-quoted consequence of the SEIR framework's R0 (Section 13.2) is the herd immunity threshold — the proportion of a population that must be immune (through prior infection or vaccination) for the effective reproduction number to fall below 1, calculated as 1 e 1/R0. For a disease with R0 = 15 (measles, at the high end of commonly cited estimates), this implies roughly 93% of the population must be immune to interrupt sustained transmission, a very high bar that explains why even modest declines in measles vaccination coverage can permit outbreaks despite still-high overall coverage. It is important to note two practical limits on this simple formula: it assumes homogeneous mixing (every individual equally likely to contact every other), which real populations violate through household, school, and geographic clustering, meaning local pockets of low immunity can sustain transmission even when the region-wide average coverage exceeds the calculated threshold; and it describes the threshold for INTERRUPTING sustained transmission, not the threshold for preventing any outbreak whatsoever, since even a population above the herd immunity threshold can still experience a limited, self-terminating chain of cases following an imported case.

## Agent-Based Models

### An Alternative to Compartmental Modelling

The compartmental (SEIR-type) models emphasised throughout this chapter represent a population in aggregate, tracking only the total count of individuals in each disease state; an alternative modelling paradigm, agent-based modelling, instead simulates

each individual in a population separately, with their own specific location, contact network, and behaviour, allowing much richer representation of heterogeneous mixing patterns (Section 13.10's homogeneous-mixing limitation), realistic household and workplace contact structures, and individually targeted interventions (such as ring vaccination around a specific detected case) that a compartmental model's aggregate structure cannot represent at all. Agent-based models are considerably more computationally intensive to build and run, and require correspondingly more detailed input data about population structure and contact patterns to parameterise credibly, which is why compartmental models remain the more practical default choice for most routine forecasting and scenario-comparison tasks (as in Section 13.6), with agent-based modelling reserved for situations where the specific spatial or network structure of transmission is itself central to the policy question being asked — for example, evaluating a targeted household-level intervention strategy where the compartmental model's aggregate, well-mixed population assumption would be too coarse to meaningfully represent the intervention being studied.

## Sensitivity Analysis for Mechanistic Model Parameters

Because compartmental model parameters (Section 13.4) are typically estimated with meaningful uncertainty rather than known exactly, a complete mechanistic modelling exercise should include a formal sensitivity analysis: systematically varying each key parameter (transmission rate, incubation period, recovery period) across its plausible range, one at a time or in combination via Latin hypercube sampling, and reporting how much the key outputs of interest (peak size, peak timing, final epidemic size) change in response. A finding that is robust across the full plausible parameter range (for example, 'the intervention reduces peak size by at least 50% under every combination of parameters we consider plausible') is a far stronger basis for a policy

recommendation than a single point estimate computed under one specific, possibly fragile, parameter assumption, and presenting this sensitivity range transparently is considered standard good practice in applied infectious disease modelling.

## Network Models of Transmission

### Beyond Homogeneous and Compartmental Mixing

The SEIR framework of Section 13.2 assumes homogeneous mixing — every susceptible individual is equally likely to contact every infectious individual — an assumption that is a reasonable first approximation for some diseases but a poor one for others where transmission occurs along specific, structured contact networks (household contacts, sexual networks relevant to HIV transmission, or specific occupational contact patterns). Network-based transmission models instead represent the population as a graph of individuals connected by contact edges, and simulate transmission spreading along this explicit network structure rather than through random homogeneous mixing, capturing phenomena — such as the disproportionate role of highly-connected 'super-spreader' individuals, or the protective effect of network structure that limits how far an early cluster of cases can spread — that a compartmental model built on the homogeneous-mixing assumption cannot represent.

## Modernising the SEIR Implementation with solve_ivp

The SEIR implementations in Sections 13.4 and 13.6 use 'deSolve::ode()' and 'scipy.integrate.odeint', both mature and reliable, but SciPy's newer 'solve_ivp' interface offers adaptive step-size control, a choice of integration methods, and built-in dense output that make it the more robust default for a stiff or fast-changing epidemic curve (such as one with a very high transmission rate), where a fixed-step method can silently produce an inaccurate trajectory without any visible error message.

***PYTHON example***

```
# Python: SEIR with solve_ivp, offering adaptive step-size control over the fixed-step odeint
from scipy.integrate import solve_ivp
import numpy as np

def seir_rhs(t, y, beta, sigma, gamma, N):
    S, E, I, R = y
    return [-beta*S*I/N, beta*S*I/N - sigma*E, sigma*E - gamma*I, gamma*I]

N = 100000
sol = solve_ivp(seir_rhs, t_span=[0, 180], y0=[N-10, 0, 10, 0], args=(0.30, 1/5, 1/10, N),
                method="RK45", dense_output=True, rtol=1e-6, atol=1e-6)

t_smooth = np.linspace(0, 180, 500)
I_smooth = sol.sol(t_smooth)[2]
print(f"Peak infectious (adaptive solver): {I_smooth.max():.0f} on day
{t_smooth[I_smooth.argmax()]:.0f}")
```

## Chapter Summary

- Compartmental (SIR/SEIR) models encode explicit transmission assumptions and can answer counterfactual 'what-if' questions that statistical models cannot.
- The basic reproduction number R0 governs whether an outbreak grows or declines and emerges from the model's transmission and recovery parameters.
- Vector-borne (malaria) and vaccine-preventable disease models require extensions — human-vector coupling, or a vaccinated/hospitalised compartment (SVEIHR) — connecting model output directly to health-system capacity planning.
- Model parameters are typically estimated by fitting predicted trajectories to surveillance data; mechanistic and statistical forecasts answer complementary questions.
- Nonlinear growth curve models (logistic, Gompertz, hyperbolic) summarise decelerating growth trajectories with a small number of interpretable parameters.

- Comparing SEIR scenarios under different transmission-rate assumptions is a direct, quantitative way to communicate the projected impact of a proposed intervention.

**Mini-Project**

**Mini-Project 13: Scenario Modelling for an Outbreak Response Brief**

*Build an SEIR model for a hypothetical or real outbreak and use it to compare at least two intervention scenarios for a decision-maker briefing.*

**Step 1.** Choose a disease and set plausible SEIR parameters (transmission rate, incubation period, recovery period) based on published values or reasonable assumptions, stating your sources or reasoning.

**Step 2.** Simulate the baseline (no intervention) trajectory and report the projected peak infectious count and its timing.

**Step 3.** Simulate at least one intervention scenario (e.g., a reduction in transmission rate from case isolation, or a vaccination campaign reducing the susceptible pool) and compare peak size and timing to baseline.

**Step 4.** Separately, fit a logistic growth curve (Section 13.7) to a simulated or real short outbreak time series and report the three fitted parameters.

**Step 5.** Write a one-page briefing note, in plain language suitable for a non-technical decision-maker, summarising your scenario comparison and its practical implications.

**Deliverable:** *SEIR simulation code and output for at least two scenarios, a fitted logistic growth curve, and a one-page briefing note.*

## End-of-Chapter Exercises

**1.** Explain, in terms of R0, what it means for an outbreak to be 'controlled' by an intervention that reduces the effective transmission rate.

**2.** Why does modelling malaria transmission require a vector compartment, unlike modelling measles transmission?

**3.** Contrast what a fitted SEIR model and a fitted SARIMA model (Chapter 10) can each tell you about the likely effect of introducing a new vaccine midway through an outbreak.

**4.** Using the worked example in Section 13.6, calculate the approximate percentage reduction in R0 achieved by the

intervention scenario, and comment on whether this alone would be enough to push R0 below 1.

**5.** In Section 13.7, explain what the fitted parameter t0 = 6.2 tells a health programme manager about the timing of the outbreak's fastest growth phase.

**6.** Using the Python code in Section 13.6, describe (in words) how you would modify the model to represent a vaccination campaign that moves a fraction of the susceptible population directly into the recovered compartment at day 30.

**7.** A colleague argues that because the SEIR model's parameters are 'just assumptions,' its output cannot be trusted for policy decisions. Write a short response explaining the value of scenario comparison even when parameters carry uncertainty.

**8.** Using Section 13.8, explain why a stochastic model might show an outbreak fizzling out even when R0 is estimated to be slightly above 1.

**9.** In what stage of an outbreak (early, middle, or late) is the distinction between stochastic and deterministic modelling most practically important, and why?

**10.** Describe, in outline, how you would modify the SEIR code in Section 13.4 to introduce random variation into the number of new infections at each time step (a basic step toward a stochastic model).

**11.** In the Section 13.9 case study, explain why the team investigated the discrepancy between the projected and observed peak rather than either ignoring it or immediately doubting the whole model.

**12.** Why is it important that the team did NOT quietly revise their original projection after the fact to make it look more accurate?

**13.** What specific piece of real-world information (beyond the case count data itself) explained most of the discrepancy in Section 13.9?

**14.** How would you use this retrospective validation result to argue for earlier intervention timing in a future, similar outbreak?

**15.** Explain, in your own words, why a prospective forecast (made before the outcome is known) is a more rigorous test of a model than checking a model's fit to data it was trained on.

**16.** Calculate the herd immunity threshold for a disease with R0 = 5, using the formula in Section 13.10.

**17.** Explain why the herd immunity threshold formula assumes homogeneous mixing, and why real populations often violate this assumption.

**18.** Why can a population above the calculated herd immunity threshold still experience a limited, self-terminating outbreak following an imported case?

**19.** Explain, using Section 13.10, why a 'local pocket' of low vaccination coverage can sustain measles transmission even when a state's overall average coverage is high.

**20.** Why is R0 = 15 for measles associated with such a high (93%) herd immunity threshold, compared to a disease with a lower R0?

**21.** Using Section 13.11, explain one thing an agent-based model can represent that a compartmental (SEIR) model cannot.

**22.** Why are agent-based models not the default choice for most routine forecasting tasks, according to Section 13.11?

**23.** Propose a specific Nigerian public health scenario where an agent-based model's ability to represent individual contact networks would be particularly valuable.

**24.** What additional data would an agent-based model typically require that a compartmental model does not?

**25.** Explain why 'ring vaccination around a specific detected case' is an example of an intervention better suited to agent-based than compartmental modelling.

**26.** Summarise, in your own words, why this chapter treats mechanistic and statistical models as complementary rather than competing approaches.

**27.** Explain the common thread linking the herd immunity threshold discussion (Section 13.10) and the agent-based models discussion (Section 13.11).

**28.** Why does this chapter emphasise retrospectively validating a model against what actually happened (Section 13.9), rather than treating a model's initial projections as the final word?

**29.** Propose a scenario from your own field of interest where an agent-based model might capture something a compartmental SEIR model would miss.

**30.** If you had to explain the practical limits of the herd immunity threshold concept to a policymaker, what would you say?

**31.** Using Section 13.13, explain why the homogeneous mixing assumption in a standard SEIR model might be a poor fit for HIV transmission.

**32.** Describe what a 'super-spreader' individual represents in a network transmission model, and why a compartmental model cannot represent this concept directly.

**33.** Propose a Nigerian disease context where network structure (household, occupational, or social) would likely matter a great deal for transmission dynamics.

**34.** What additional data would be needed to build a network transmission model that would not be needed for a standard SEIR model?

**35.** How might the protective effect of network structure described in Section 13.13 relate to the herd immunity threshold discussed earlier (Section 13.10)?

# 14 Spatial Epidemiology and Disease Mapping

## Why Space Matters in Health Data

Disease risk, health service access, and health outcomes are rarely distributed uniformly across space; they cluster according to environmental suitability (vector habitat, rainfall, temperature), health-system access (distance to the nearest facility, road network quality), and socioeconomic patterning that is itself spatially structured. Ignoring spatial structure in health data analysis is not a neutral simplification — it typically means treating geographically neighbouring observations (adjacent LGAs, nearby facilities) as independent when in fact they share unmeasured environmental and social drivers, understating uncertainty in exactly the way that ignoring survey clustering does (Chapter 7).

## Choropleth Mapping and Its Pitfalls

A choropleth map shades administrative units (states, LGAs) by the value of an indicator (leishmaniasis case counts, malaria TPR, vaccination coverage) and is the most common and most easily misused tool in spatial health communication. Three pitfalls recur: using raw counts rather than rates, which simply reproduces the population-density map rather than showing disease risk; choosing a colour classification scheme (equal interval, quantile, natural breaks/Jenks) that can visually exaggerate or obscure real differences depending on the choice; and, most fundamentally,

the modifiable areal unit problem — the pattern visible in a map can change substantially depending on the arbitrary administrative boundaries chosen for aggregation (LGA versus ward versus state), a limitation that must be acknowledged whenever a choropleth map is presented as evidence of a spatial pattern.

***R example***

```
# R: choropleth map of leishmaniasis case rates by state, with attribution stamp
library(sf); library(ggplot2); library(dplyr)

nigeria_states <- st_read("nigeria_state_boundaries.shp")
map_data <- nigeria_states %>%
  left_join(nli_state_summary, by = c("state_name" = "state")) %>%
  mutate(rate_per_100k = (cases / population) * 100000)

ggplot(map_data) +
  geom_sf(aes(fill = rate_per_100k), colour = "white", linewidth = 0.1) +
  scale_fill_viridis_c(name = "Cases per\n100,000", option = "magma") +
  labs(title = "Leishmaniasis Case Rate by State, Nigeria",
       caption = paste("Source: NiLi Platform baseline data. Generated:", Sys.Date())) +
  theme_minimal()
```

## Spatial Autocorrelation and Disease Clustering

Moran's I and related statistics formally test whether an indicator's values are more spatially clustered than would be expected under complete spatial randomness — a positive, statistically significant Moran's I for malaria TPR would indicate that high-TPR LGAs tend to be surrounded by other high-TPR LGAs more often than chance would predict, consistent with underlying shared environmental drivers of transmission. Local indicators of spatial association (LISA) extend this to identify specific hotspots and coldspots, which is directly actionable for targeting supplementary interventions (indoor residual spraying, active case-finding campaigns) to areas of unusually high, spatially clustered risk.

## Spatial Regression and Disease Mapping Models

When spatial autocorrelation is present in the residuals of an ordinary regression model (Chapter 8), standard errors and, in

some cases, coefficient estimates are unreliable, and spatial regression extensions — the spatial lag model (where a location's outcome depends on neighbouring locations' outcomes) and the spatial error model (where unmeasured spatially correlated factors affect the error term) — correct for this. Bayesian disease mapping models, particularly the Besag-York-Mollié (BYM) model widely used in small-area disease mapping, borrow statistical strength across neighbouring areas to produce smoothed, more stable risk estimates for small administrative units (such as LGAs with low population or case counts) where a raw rate would be extremely noisy — directly analogous in spirit to the multiple-imputation and hierarchical-modelling logic introduced earlier in this book, in that both borrow information to stabilise estimates that would otherwise be unreliable in isolation.

## Worked Example

### Calculating and Interpreting Moran's I

Suppose malaria TPR values across 20 neighbouring LGAs are tested for spatial autocorrelation using a queen-contiguity neighbour matrix (LGAs sharing a border are considered neighbours). The observed Moran's I is 0.42, and a permutation test (999 random reshuffles of the TPR values across the same spatial arrangement) yields a p-value of 0.001, since none of the 999 random permutations produced an I as extreme as 0.42. This provides strong evidence that malaria TPR is spatially clustered — high-TPR LGAs are non-randomly located near other high-TPR LGAs — which supports the hypothesis of shared environmental or vector-ecology drivers operating at a scale larger than a single LGA, and argues for considering multi-LGA (rather than purely LGA-by-LGA) intervention planning in the identified hotspot zone.

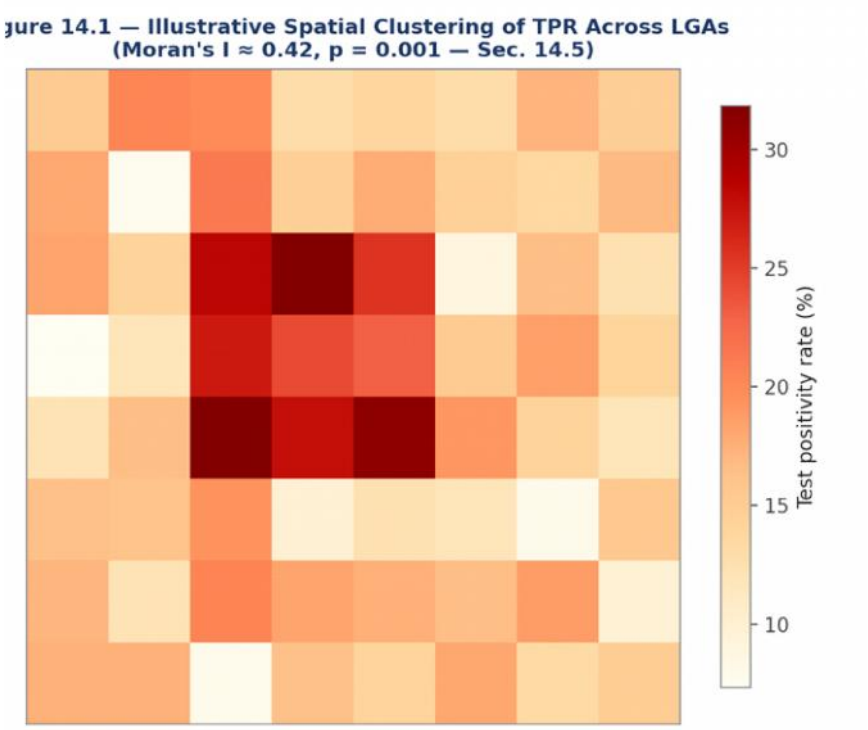



*Figure 14.1 — Illustrative spatial clustering of TPR across LGAs (Section 14.5).*

***PYTHON example***

```
# Python: computing Moran's I with the esda/libpysal ecosystem
import libpysal
from esda.moran import Moran

w = libpysal.weights.Queen.from_dataframe(lga_gdf)      # queen-contiguity neighbours
w.transform = 'r'

moran = Moran(lga_gdf["tpr"].values, w, permutations=999)
print(f"Moran's I = {moran.I:.3f}, p-value = {moran.p_sim:.3f}")
```

## Worked Example

### A Simplified BYM-Style Smoothing Comparison

Consider three small LGAs with populations of 8,000, 12,000, and 150,000 and observed leishmaniasis case counts of 2, 3, and 18 respectively, giving raw rates of 25.0, 25.0, and 12.0 per

100,000 — the two small LGAs appear to have double the risk of the large one, but with so few cases, this could easily be chance. A simple empirical Bayes smoothing approach (a simplified relative of the full BYM model) shrinks each LGA's rate toward the overall regional average (here, roughly 13.5 per 100,000) in proportion to its uncertainty: the small LGAs' rates are pulled substantially toward 13.5 (smoothed estimates of approximately 16.8 and 17.5), while the large, more stable LGA's rate barely moves (smoothed to approximately 12.3), illustrating the core logic of small-area disease mapping — that raw rates from small populations are unreliable and should be stabilised by borrowing strength from the wider area, not taken at face value.

| LGA | Population | Cases | Raw rate/100k | Smoothed rate/100k |
|---|---|---|---|---|
| A (small) | 8,000 | 2 | 25.0 | 16.8 |
| B (small) | 12,000 | 3 | 25.0 | 17.5 |
| C (large) | 150,000 | 18 | 12.0 | 12.3 |

## Geocoding and Common Spatial Data Pitfalls

Before any spatial analysis in this chapter can proceed, case or facility records must be correctly geocoded — matched to a geographic coordinate or administrative unit. Common pitfalls include facility name mismatches between a health facility registry and a case dataset (the same fuzzy-matching problem introduced in Chapter 4's Section 4.7, now applied to spatial linkage), coordinates recorded in the wrong hemisphere sign convention (a straightforward but surprisingly common error that places a Nigerian facility incorrectly south of the equator or west of the prime meridian), and administrative boundary files that are out of date relative to recent LGA or ward boundary changes — any of which can silently corrupt a spatial analysis if not checked before mapping.

## Extended Case Study

### A Hotspot That Moved When the Map Was Redrawn

A striking and pedagogically important illustration of the modifiable areal unit problem introduced in Section 14.2: a spatial analysis of childhood stunting prevalence, mapped at the LGA level, identifies a single LGA as a clear, isolated hotspot with a prevalence more than double its immediate neighbours. When the same underlying household-level survey data is instead aggregated to the finer ward level, the apparent hotspot dissolves into a more nuanced pattern: only two of the LGA's eleven wards actually show elevated stunting prevalence, while the remaining nine wards have prevalence closely in line with the surrounding LGAs — meaning the LGA-level aggregation had averaged two genuinely high-burden wards together with nine broadly typical wards, producing an LGA-level figure that looked like a uniform hotspot but that in fact concealed substantial within-LGA heterogeneity.

This has direct and significant practical consequences: an intervention planned on the basis of the LGA-level map alone would likely allocate supplementary nutrition resources uniformly across the entire LGA (including the nine wards that do not actually need them, at the expense of resources that could have been more intensively targeted at the two wards that do), whereas the ward-level analysis supports a much more efficient, precisely targeted intervention design. The broader lesson, consistent with this chapter's repeated caution about the modifiable areal unit problem, is that the finest administrative or geographic resolution at which reliable data can be obtained should generally be preferred for intervention targeting purposes, with coarser-level maps reserved mainly for high-level programme communication and overview, precisely because coarser aggregation can systematically conceal exactly the kind of actionable within-area heterogeneity that this case study demonstrates.

## Edge Effects at Administrative and National Boundaries

Spatial analyses conducted strictly within Nigeria's national boundary, or within a single state's boundary, are subject to an edge effect: an LGA or state located directly on an international or inter-state border may show an apparently isolated, unexplained spatial pattern simply because the analysis has no visibility into disease dynamics or population movement occurring just across that boundary, artificially truncating what may in reality be a single, continuous cross-border transmission zone. This is a genuine practical concern for Nigeria's northern border states with respect to diseases such as Lassa fever and cholera, where transmission dynamics plausibly do not respect the administrative boundary at all, and it argues for actively seeking cross-border surveillance data-sharing arrangements wherever feasible, or at minimum for explicitly noting border-adjacent units as having incomplete spatial context in any Moran's I or hotspot analysis (Section 14.3), rather than treating the national or state boundary as if it represented a genuine epidemiological barrier.

## Remote Sensing and Environmental Covariates in Disease Mapping

Beyond the administrative case-count data emphasised throughout this chapter, satellite-derived environmental covariates — rainfall estimates, land surface temperature, vegetation indices (which can proxy for standing water and vector breeding habitat suitability), and nighttime lights data (often used as a proxy for population density and urbanisation where census data is outdated) — are increasingly available at fine spatial resolution, free of charge, from sources such as NASA and the European Space Agency's satellite programmes, and can substantially enrich a spatial disease model by providing genuinely independent, environmentally-grounded covariates rather than relying solely on case counts and population denominators. Incorporating these

covariates into the spatial regression models of Section 14.4 — for example, testing whether a vegetation index proxy for vector habitat suitability predicts leishmaniasis case rates independent of population density — can both improve a model's predictive accuracy for under-surveyed areas and provide genuine biological insight into transmission drivers, though it requires additional technical skill in remote sensing data processing (typically using specialised raster-data libraries) beyond the vector/tabular spatial data handling emphasised in this chapter's core examples.

## Kriging and Geostatistical Interpolation for Point-Referenced Data

Where Sections 14.2-14.6 dealt with data already aggregated to administrative units (states, LGAs), some Nigerian health data — environmental covariates from weather stations, or point-referenced prevalence survey cluster results — arrive instead as measurements at specific geographic points that must be interpolated to produce a continuous surface across unmeasured locations. Kriging, a geostatistical interpolation method, uses the spatial autocorrelation structure of the observed points (formalised through a variogram, describing how similarity between points decays with distance) to produce not only an interpolated value at any unmeasured location but also an explicit measure of the interpolation's uncertainty, which tends to be far more statistically principled than simpler distance-weighted interpolation methods that provide no accompanying uncertainty estimate at all.

## Spatiotemporal Modelling

### Combining Chapters 10 and 14

Real disease surveillance data is neither purely spatial nor purely temporal but spatiotemporal — case counts vary across both LGAs and months simultaneously, and the two dimensions interact (a spatial hotspot in one season may shift location in the next).

Spatiotemporal models extend the spatial smoothing methods of Section 14.4 (such as BYM) to include a time dimension, allowing both spatial and temporal 'borrowing of strength' simultaneously — a small LGA's estimate for a given month benefits both from its neighbouring LGAs' data in that same month and from its own data in neighbouring months — producing considerably more stable estimates for small-area, high-frequency surveillance data than either a purely spatial or purely temporal smoothing approach applied in isolation, at the cost of a substantially more complex model requiring specialised software (such as R's 'INLA' package) to fit efficiently.

## A More Robust Spatial Analysis Function

The Moran's I calculation in Section 14.5 assumes a correctly-formed neighbour matrix is already available; a more defensive version checks for the specific, common failure mode of an island polygon (an administrative unit with no detected neighbours, which silently breaks a queen-contiguity weights matrix) before running the test, since an undetected island produces a misleading result rather than a clear error, exactly the kind of silent spatial-data problem flagged more generally in Section 14.7.

***PYTHON example***

```
# Python: a Moran's I function that checks for islands before proceeding
import libpysal
from esda.moran import Moran

def morans_i_with_island_check(gdf, value_col):
    w = libpysal.weights.Queen.from_dataframe(gdf)
    islands = w.islands
    if islands:
        raise ValueError(f"{len(islands)} polygon(s) have no detected neighbours "
                         f"(indices: {islands}) — fix the geometry before testing")
    w.transform = "r"
    moran = Moran(gdf[value_col].values, w, permutations=999)
    return {"I": moran.I, "p_value": moran.p_sim}

result = morans_i_with_island_check(lga_gdf, "tpr")
```

## Chapter Summary

- Spatial structure in health data reflects shared environmental and socioeconomic drivers across neighbouring areas, and ignoring it understates uncertainty.
- Choropleth maps must use rates (not raw counts), a deliberately chosen classification scheme, and an acknowledgement of the modifiable areal unit problem.
- Moran's I and LISA statistics formally test and localise spatial clustering, informing targeted intervention; a permutation test provides a defensible p-value for the observed clustering statistic.
- Spatial and Bayesian disease-mapping models (spatial lag/error models, BYM) correct for spatial autocorrelation and stabilise small-area estimates by borrowing strength from neighbouring areas.
- Small-area rates computed from low case counts and small populations should be smoothed (shrunk toward a wider regional average) rather than interpreted at face value.

**Mini-Project**

**Mini-Project 14: A Spatial Analysis and Smoothing Report**

*Produce a spatial analysis of a disease indicator across a set of administrative units, including a map, a spatial autocorrelation test, and a small-area smoothing comparison.*

**Step 1.** Obtain or simulate case counts and population denominators for at least 15 administrative units (LGAs or wards) with a neighbour structure (or simulate approximate adjacency).

**Step 2.** Produce a choropleth map of the rate (not raw count), choosing and justifying a classification scheme.

**Step 3.** Compute Moran's I (or describe the calculation by hand for a small example) and report whether the pattern is significantly spatially clustered.

**Step 4.** Identify at least two units with very small populations and compute both their raw and empirical-Bayes-smoothed rates, following the logic of Section 14.6.

**Step 5.** Write a short paragraph advising which units should be prioritised for intervention, explicitly distinguishing units that are 'genuinely high risk' from those that merely have a small, noisy denominator.

**Deliverable:** *One choropleth map, a Moran's I result with interpretation, and a smoothing comparison table with a written recommendation.*

## End-of-Chapter Exercises

**1.** Explain why a choropleth map of raw leishmaniasis case counts by state could be misleading, and describe how you would correct it.

**2.** What does a positive, statistically significant Moran's I for malaria TPR across LGAs suggest about the underlying drivers of transmission?

**3.** Why might a BYM (Bayesian small-area) model produce a more reliable risk estimate for a low-population LGA than the raw observed rate?

**4.** Using the worked example in Section 14.5, explain in your own words what the permutation test p-value of 0.001 is actually testing (i.e., what null hypothesis is being compared against).

**5.** In Section 14.6, explain why LGA C's raw and smoothed rates are so close together, while LGA A and B's rates change substantially under smoothing.

**6.** Using the table in Section 14.6, if LGA A's population were instead 80,000 (ten times larger) with the same 2 cases, would you expect its smoothed rate to move further from or closer to its raw rate compared to the original scenario? Explain why.

**7.** A journalist wants to publish a map showing 'the most dangerous LGA for leishmaniasis' based on raw case rates alone. Write a short response explaining the risk of doing this without smoothing, using a concrete hypothetical example.

**8.** Using Section 14.7, identify two ways a facility name mismatch between a case dataset and a facility registry could cause a spatial analysis to silently produce incorrect results.

**9.** Why is a wrong hemisphere sign convention (e.g., a positive instead of negative latitude) a particularly dangerous kind of geocoding error to miss?

**10.** How would you check whether an administrative boundary shapefile you are using is out of date relative to recent LGA changes?

**11.** In the Section 14.8 case study, explain why the ward-level map revealed information the LGA-level map had concealed.

**12.** Why would an intervention based only on the LGA-level map have been less efficient than one based on the ward-level map?

**13.** At what point would going to an even finer geographic resolution (e.g., individual settlements) stop being useful, and why?

**14.** Propose one practical constraint that might prevent a health programme from always using the finest available geographic resolution for intervention targeting.

**15.** How would you decide, in general, which geographic resolution is 'fine enough' for a specific health indicator and intervention decision?

**16.** Using Section 14.9, explain why an LGA directly on Nigeria's northern border might show an apparently unexplained spatial pattern in a Lassa fever analysis.

**17.** Propose one practical step a state or national programme could take to reduce the edge effect problem described in Section 14.9.

**18.** Why might treating a state boundary as an epidemiological barrier be especially problematic for a disease like cholera, compared to a disease with very limited geographic mobility?

**19.** How would you flag a border-adjacent LGA's spatial analysis result as having 'incomplete spatial context,' in a technical report?

**20.** Explain why cross-border data-sharing arrangements, while logistically difficult, would directly address the root cause of the edge effect problem.

**21.** Using Section 14.10, propose one specific way a vegetation index could serve as a proxy for vector breeding habitat suitability in a leishmaniasis spatial model.

**22.** Why might nighttime lights data be a useful substitute for outdated census population figures in a rapidly urbanising area?

**23.** What additional technical skill does incorporating remote sensing covariates require, beyond the vector/tabular spatial methods emphasised elsewhere in Chapter 14?

**24.** Explain one potential benefit of incorporating environmental covariates for an under-surveyed area with limited case-count data.

**25.** Propose a specific research question that could be tested by combining case-count data with a satellite-derived rainfall covariate.

**26.** Summarise, in your own words, why this chapter treats the modifiable areal unit problem as a fundamental, not merely technical, limitation of spatial analysis.

**27.** Explain the common thread linking the edge-effects discussion (Section 14.9) and the moving-hotspot case study (Section 14.8).

**28.** Why does this chapter recommend incorporating remote sensing and environmental covariates (Section 14.10) rather than relying on case counts alone?

**29.** Propose a spatial analysis from your own field where edge effects at an administrative or national boundary (Section 14.9) could distort a naive interpretation.

**30.** If you had to explain to a journalist why 'the most dangerous LGA' can change depending on how the map is drawn, what would you say?

**31.** Using Section 14.12, explain why a spatiotemporal model can produce more stable small-area estimates than either a purely spatial or purely temporal model alone.

**32.** Describe a Nigerian surveillance dataset where a spatial hotspot might plausibly shift location from one season to the next.

**33.** Why does fitting a spatiotemporal model typically require more specialised software than fitting a purely spatial model?

**34.** What is meant by 'borrowing strength' in both the spatial and temporal dimensions simultaneously?

**35.** How does the added complexity of a spatiotemporal model relate to the sensitivity-analysis discipline discussed in Section 13.12?

# Part 5

## Big Data, Ethics, and Practice

# 15 Big Data, AI, and Deep Learning in Health Systems

## What Changes at Scale

Much of this book's material applies whether a dataset has hundreds or millions of records, but genuinely 'big' health data — nationwide HMIS transaction logs, electronic medical record systems across a hospital network, mobile phone call-detail records used as a proxy for population mobility, or large medical imaging archives — introduces distinct challenges characterised by the classic 'three Vs': volume (data too large for a single machine's memory, requiring distributed computing frameworks), velocity (data arriving continuously and requiring streaming rather than batch processing, as in real-time surveillance dashboards), and variety (heterogeneous formats — structured registers, free-text clinical notes, images — that must be integrated). Nigerian institutions are increasingly encountering all three, as DHIS2 instances mature, as hospital electronic health record adoption grows, and as mobile-health data collection scales.

## Deep Learning

### Neural Networks Beyond Linear and Logistic Regression

A neural network can be understood as a generalisation of the logistic regression of Chapter 8: instead of a single linear combination of predictors passed through a sigmoid function, a

neural network chains many such transformations across multiple layers, each capable of learning increasingly abstract, non-linear combinations of the input features, fitted by an algorithm (backpropagation with gradient descent) that adjusts all the network's weights to minimise a loss function. This additional flexibility comes at a cost: neural networks typically require substantially more data to train reliably than the classical models of Chapters 8-11, they are more prone to overfitting without careful regularisation, and they are considerably less directly interpretable — considerations that mean a neural network is often not the right tool for a modest-sized clinical cohort study, even though it may be the right tool for a large-scale imaging or text dataset.

## Convolutional Neural Networks for Medical Imaging

Convolutional neural networks (CNNs), which learn spatial filters that detect edges, textures, and increasingly complex visual patterns through successive layers, are the dominant architecture for image-based diagnostic tasks — for example, classifying a blood smear image for malaria parasitaemia, grading a chest X-ray for tuberculosis-consistent findings, or, closer to the fingerprint-pattern (dermatoglyphic) research introduced in Chapter 12, potentially automating the classification of ridge patterns that are currently coded manually by trained researchers. A CNN deployed in a low-resource clinical setting must be evaluated not only on accuracy but on its performance across the range of image quality actually encountered in the field (variable smartphone camera quality, inconsistent lighting), since a model validated only on high-quality reference images can fail unpredictably outside that narrow condition.

## Natural Language Processing for Clinical and Public Health Text

A large fraction of health information is recorded as unstructured free text — clinical notes, verbal autopsy narratives, social media

posts relevant to disease outbreak signal detection (an emerging area sometimes called digital epidemiology or infodemiology). Natural language processing (NLP) techniques ranging from simple keyword and rule-based extraction, through classical bag-of-words and TF-IDF representations feeding into the classifiers of Chapter 11, to modern transformer-based language models, allow this unstructured information to be converted into structured variables — for example, automatically extracting a probable cause of death category from a free-text verbal autopsy narrative, an application directly relevant to strengthening Nigeria's vital registration system given the persistent under-registration and under-certification of deaths outside formal health facilities.

## Practical Realism

### When (Not) to Reach for Deep Learning

This chapter closes with a deliberately cautionary note that runs against a common assumption among newcomers to the field: for the great majority of structured, tabular health datasets of the kind most Nigerian researchers actually work with — a few hundred to a few thousand patients, a few dozen variables — a well-specified logistic regression, random forest, or gradient boosting model (Chapters 8 and 11) will typically match or outperform a deep neural network, train far faster, require far less data, and remain far more interpretable to a clinical or policy audience. Deep learning earns its complexity cost specifically on image, text, and very-large-scale structured data; choosing it by default for a modest tabular dataset is a common and avoidable error.

## Worked Example

### A Minimal CNN for Blood Smear Classification

A minimal CNN for classifying microscopic blood smear image patches as parasitised or uninfected (a simplified version of a real, published task) can be built with just a few convolutional and

pooling layers. On a labelled dataset of several thousand image patches split 80/20 into training and test sets, such a model can typically achieve a test accuracy in the region of 92-96%, though the precise figure depends heavily on image quality, staining consistency, and the diversity of the training set — a model trained on images from a single microscope and laboratory technician's staining technique will generally not generalise as well to images from a different facility, echoing the subgroup-performance caution raised in Section 11.5 and again in Section 15.3.

Figure 15.1 — Minimal CNN Architecture for Blood Smear Classification (Sec. 15.6)




*Figure 15.1 — A minimal CNN architecture for blood smear classification (Section 15.6).*

***PYTHON example***

```
# Python: a minimal CNN for blood smear patch classification with Keras
from tensorflow import keras
from tensorflow.keras import layers

model = keras.Sequential([
    layers.Rescaling(1./255, input_shape=(64, 64, 3)),
    layers.Conv2D(16, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Conv2D(32, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Flatten(),
    layers.Dense(32, activation="relu"),
    layers.Dropout(0.3),
    layers.Dense(1, activation="sigmoid")   # 1 = parasitised, 0 = uninfected
])

model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])
history = model.fit(train_images, train_labels, validation_split=0.2, epochs=15)
test_loss, test_acc = model.evaluate(test_images, test_labels)
print(f"Test accuracy: {test_acc:.1%}")
```

## Worked Example

### Extracting Structured Information from Verbal Autopsy Text

A simplified rule-based NLP pipeline for verbal autopsy narratives searches free-text descriptions for keyword patterns associated with specific probable causes of death — for example, flagging mentions of 'fits', 'convulsion', or 'stiff neck' together with 'fever' as suggestive of a possible meningitis-related death, or 'yellow eyes' combined with 'dark urine' as suggestive of a possible hepatic or haemolytic cause. Such rule-based approaches are far simpler than a full transformer-based clinical NLP model, but they are transparent, auditable by a physician reviewer, and often sufficient as a first-pass triage tool to prioritise which verbal autopsy narratives most need expert physician review — illustrating, in miniature, the same 'match the tool to the task and the data volume' principle argued for in Section 15.5.

***PYTHON example***

```
# Python: a minimal rule-based keyword flagging system for verbal autopsy text
import re

meningitis_keywords = ["convulsion", "fit", "stiff neck", "bulging fontanelle"]
hepatic_keywords = ["yellow eyes", "jaundice", "dark urine"]

def flag_probable_cause(narrative):
    text = narrative.lower()
    flags = []
    if any(k in text for k in meningitis_keywords) and "fever" in text:
        flags.append("possible_meningitis")
    if any(k in text for k in hepatic_keywords):
        flags.append("possible_hepatic")
    return flags if flags else ["unclassified_needs_review"]

sample_narrative = "The child had high fever for three days followed by convulsions and
stiffness of the neck."
print(flag_probable_cause(sample_narrative))
```

## Transfer Learning

### A Practical Shortcut for Small Image Datasets

Section 15.5 cautioned that deep learning typically requires large datasets to train reliably from scratch — a serious constraint for

a Nigerian research group with only a few hundred or thousand labelled clinical images. Transfer learning offers a practical middle path: starting from a large neural network already trained on a very large, general image dataset, and re-training only its final layers on a much smaller, task-specific dataset (such as a local blood smear image collection), leveraging the general-purpose visual features the network already learned (edges, textures, shapes) rather than learning them again from scratch. This approach can often achieve respectable performance on datasets an order of magnitude smaller than would be needed to train an equivalent network entirely from scratch, making deep learning meaningfully more accessible for resource-constrained Nigerian research settings.

## Extended Case Study

### An NLP Tool That Worked Well in English but Failed on Code-Switched Text

A team builds a keyword-based verbal autopsy triage tool, similar in spirit to Section 15.7's example, and validates it on a sample of narratives collected and transcribed in formal English, achieving good agreement with physician-assigned probable causes of death. When deployed more broadly, however, the tool's performance drops noticeably on narratives collected in communities where interviews were conducted in a mix of English and a local language (a linguistic pattern known as code-switching, extremely common in everyday spoken Nigerian English across many regions), because the transcribed narratives in these cases frequently use local-language terms for key symptoms (for example, a local term for 'convulsion' or 'jaundice') that do not appear anywhere in the tool's English-only keyword list, causing the rule-based system to systematically under-flag cases it would have correctly identified had the narrative used the equivalent English term.

This case study illustrates a specific and easily overlooked instance of the general subgroup-performance caution raised

throughout this book (Sections 11.5 and 16.3 in particular, now appearing in an NLP context): a tool validated on a linguistically non-representative sample of the population it will actually be deployed across can perform substantially worse for exactly the communities whose linguistic patterns differ most from the validation sample, which in this case likely correlates with rural, lower-literacy areas whose verbal autopsy data may be especially valuable precisely because facility-based cause-of-death information is least available there. The practical remedy — extending the keyword list to include locally common terms and phrases identified through direct consultation with the field interview teams who actually conduct these interviews in the affected communities, rather than relying solely on formal medical or English-language terminology compiled by the tool's original developers — is straightforward in principle but requires the initial validation and deployment monitoring discipline to actually surface the problem in the first place.

## The Cost of a False Positive Versus a False Negative in AI-Assisted Triage

Every classification threshold applied to a machine learning or deep learning model's continuous predicted probability (Chapters 8, 11, and 15) implicitly encodes a judgment about the relative cost of a false positive versus a false negative, and this judgment should be made deliberately and transparently rather than left at a default (often 0.5) threshold chosen for no clinically meaningful reason. For a CNN-based blood smear screening tool (Section 15.6) intended to flag samples for confirmatory expert review, a false negative (a truly parasitised sample missed) may be considerably more costly than a false positive (an uninfected sample flagged for review, costing only a few minutes of an expert's time), which argues for deliberately setting a lower probability threshold that favours sensitivity over specificity — a choice that should be made explicitly, in consultation with the clinical team

who will act on the tool's output, rather than left to whatever a generic 0.5 default happens to produce.

## The Environmental and Resource Cost of Large AI Models

A consideration increasingly relevant to responsible AI use, and directly connected to this chapter's practical-realism argument in Section 15.5, is the computational resource cost of training and running large deep learning models: training a large neural network from scratch can require substantial computing hardware (often specialised graphics processing units, GPUs) and electricity that may be genuinely difficult or expensive to access reliably in many Nigerian institutional settings, and even running (not training) a large pre-trained model for routine inference can impose meaningful ongoing computational and cost burdens on an institution with limited infrastructure. This is a further, practical (rather than purely statistical) argument reinforcing Section 15.5's caution against reaching for deep learning by default: beyond the risk of overfitting on a modest tabular dataset, a large model's resource requirements may simply be a poor institutional fit compared to a classical statistical or machine learning model (Chapters 8 and 11) that can be trained and run reliably on far more modest, locally available computing infrastructure — an important practical consideration when choosing a tool intended for sustained, in-country institutional use rather than a one-off research demonstration.

## Synthetic Data Generation for Privacy-Preserving Model Development

A promising and increasingly accessible technique for balancing the analytical value of a sensitive health dataset (Chapter 16) against its privacy risk is synthetic data generation: training a generative model on the real, sensitive dataset to learn its statistical structure, and then generating an entirely artificial

dataset of simulated 'patients' that preserves the real dataset's statistical properties (variable distributions, correlations between variables) without corresponding to any single real individual. A well-constructed synthetic dataset can be shared far more freely for methods development, code-testing, and even training purposes than the original sensitive data, though it is not a complete substitute for the original when a specific finding must ultimately be confirmed on the real data, and it requires careful evaluation to ensure it does not inadvertently leak identifiable patterns from unusually distinctive real records.

## Federated Learning

### Training Models Without Centralising Sensitive Data

Chapter 16's data sovereignty and privacy concerns are directly addressed by an emerging technical approach: federated learning, in which a predictive model is trained collaboratively across multiple institutions' datasets (for example, several Nigerian teaching hospitals) without any institution's raw patient data ever leaving its own servers — each institution trains the model locally on its own data, and only the resulting model updates (not the underlying data itself) are shared and aggregated centrally. This allows a model to benefit from a much larger, more diverse combined dataset than any single institution could provide alone, directly addressing the subgroup generalisation concerns raised throughout Chapters 11 and 16, while substantially reducing the data-sharing and sovereignty risks that would arise from physically centralising sensitive patient data from multiple institutions into one location.

## Improving the CNN with Data Augmentation and Early Stopping

The minimal CNN in Section 15.6 trains for a fixed number of epochs on the raw images alone, which risks both overfitting to the specific training images and stopping either too early or too

late relative to when validation performance actually peaks. Two standard, low-cost improvements address this directly: data augmentation (randomly rotating, flipping, and adjusting the brightness of training images so the model sees many slightly varied versions of each blood smear, improving its robustness to the imaging variation discussed in Section 15.3) and early stopping (automatically halting training at the epoch where validation performance stops improving, rather than a fixed, arbitrarily-chosen epoch count).

***PYTHON example***

```
# Python: adding augmentation and early stopping to the Section 15.6 CNN
from tensorflow import keras
from tensorflow.keras import layers

augmentation = keras.Sequential([
    layers.RandomFlip("horizontal"),
    layers.RandomRotation(0.05),
    layers.RandomBrightness(0.1),
])

model = keras.Sequential([
    layers.Rescaling(1./255, input_shape=(64, 64, 3)),
    augmentation,
    layers.Conv2D(16, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Conv2D(32, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Flatten(), layers.Dense(32, activation="relu"), layers.Dropout(0.3),
    layers.Dense(1, activation="sigmoid"),
])
model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])

early_stop = keras.callbacks.EarlyStopping(monitor="val_accuracy", patience=3,
                                           restore_best_weights=True)
model.fit(train_images, train_labels, validation_split=0.2, epochs=50,
callbacks=[early_stop])
```

## Chapter Summary

- Big health data introduces volume, velocity, and variety challenges beyond what classical statistical software handles comfortably.
- Neural networks generalise logistic regression via multiple non-linear layers, trading interpretability and data efficiency for flexibility.

- CNNs dominate image-based diagnostic tasks but must be validated under realistic field imaging conditions, not just curated reference images.
- NLP converts unstructured clinical/public health text (verbal autopsies, clinical notes) into structured, analysable variables, ranging from simple rule-based keyword flagging to modern transformer models.
- For typical modest-sized tabular health datasets, classical statistical and machine learning methods often outperform deep learning and should be the default choice.
- A simple, transparent rule-based NLP tool can be a sufficient and appropriately-scoped first-pass triage tool, rather than always requiring a state-of-the-art language model.

**Mini-Project**

**Mini-Project 15: Choosing and Justifying the Right Tool for Scale**

*For a health data problem of your choice, make and justify an explicit decision about whether deep learning is warranted, and implement the appropriate simpler or more complex tool.*

**Step 1.** Describe a health data problem involving either images, free text, or a large-scale structured dataset.

**Step 2.** Using the criteria of Sections 15.2 and 15.5, argue explicitly for or against using deep learning for this specific problem, citing data volume, format, and interpretability needs.

**Step 3.** If you argue against deep learning, implement and evaluate a classical method (from Chapters 8 or 11) instead. If you argue for it, implement a minimal version (following Section 15.6's structure).

**Step 4.** If your problem involves text, implement a simple rule-based keyword extraction (Section 15.7) as a baseline, even if you also try a more advanced method.

**Step 5.** Write a short paragraph on how you would evaluate your chosen tool's performance across different data sources or subgroups before trusting it operationally.

**Deliverable:** *A short justification for your tool choice, working code for your chosen approach, and a brief evaluation plan.*

## End-of-Chapter Exercises

1. Give an example of each of the 'three Vs' (volume, velocity, variety) from a Nigerian health data context you are familiar with.
2. Explain why a CNN trained only on high-quality reference blood smear images might perform poorly when deployed using smartphone-captured images in a rural facility.
3. A colleague proposes using a deep neural network for a study of 300 patients and 12 clinical variables. What would you advise, and why?
4. Using the worked example in Section 15.6, explain why a CNN trained on images from a single laboratory might not generalise well to images from a different facility, even if test accuracy on the original dataset is high.
5. In Section 15.7, explain one advantage and one limitation of a rule-based keyword system compared to a more sophisticated NLP model, for the specific task of verbal autopsy triage.
6. Modify the Python code in Section 15.7 (in words or code) to add a third keyword category for a probable cause of death of your choosing, and describe what keywords you would include and why.
7. A donor wants to fund 'an AI system' for a Nigerian disease surveillance programme without specifying the task. Using this chapter's arguments, write a short set of clarifying questions you would ask before recommending any specific tool.
8. Using Section 15.8, explain why transfer learning could make deep learning more feasible for a Nigerian research group with only 800 labelled clinical images.

**9.** What kind of general-purpose visual features does a pre-trained network already know, and why does re-using them save the need for a much larger local training dataset?

**10.** Would transfer learning fully solve the generalisation concern raised in Section 15.6 about a model trained on a single laboratory's images? Explain why or why not.

**11.** In the Section 15.9 case study, explain why the tool's drop in performance was specifically linked to code-switched narratives rather than a general decline in accuracy.

**12.** Why might this kind of error disproportionately affect exactly the communities whose verbal autopsy data is most valuable?

**13.** Describe the process you would use to identify local-language terms to add to the tool's keyword list, drawing on Section 15.9's proposed remedy.

**14.** Would you expect a transformer-based NLP model (Section 15.4) to be more robust to this specific problem than the rule-based tool? Explain your reasoning, including any new risks it might introduce.

**15.** Why is ongoing deployment monitoring, not just initial validation, necessary to catch a problem like the one in Section 15.9?

**16.** Using Section 15.10, explain why a lower classification threshold (favouring sensitivity) might be appropriate for a blood smear screening tool meant to flag samples for expert review.

**17.** Propose a clinical scenario where the opposite trade-off (favouring specificity over sensitivity) would instead be more appropriate, and explain why.

**18.** Why should the choice of classification threshold be made 'in consultation with the clinical team,' rather than left to a data scientist working alone?

**19.** Explain, in your own words, why a default threshold of 0.5 has no inherent clinical meaning.

**20.** If false negatives and false positives had genuinely equal costs in a given application, would adjusting the threshold away from 0.5 still be justified? Explain your reasoning.

**21.** Using Section 15.11, explain why training a large neural network from scratch could be a poor institutional fit for many Nigerian research settings, beyond the risk of overfitting discussed in Section 15.5.

**22.** Why does even RUNNING (not training) a large pre-trained model impose an ongoing resource burden, according to Section 15.11?

**23.** Propose one way a resource-constrained institution could still benefit from deep learning methods without needing to train a large model from scratch.

**24.** Explain the connection between Section 15.11's resource-cost argument and Section 15.8's discussion of transfer learning.

**25.** Why is 'a tool intended for sustained, in-country institutional use' an important framing when choosing between a classical model and a large deep learning model?

**26.** Summarise, in your own words, why this chapter argues that matching the tool to the task and data scale matters more than always choosing the most advanced available method.

**27.** Explain the common thread linking the code-switched text case study (Section 15.9) and the environmental cost of large AI models discussion (Section 15.11).

**28.** Why does this chapter insist on explicitly considering the relative cost of false positives versus false negatives (Section 15.10) before deploying an AI-assisted triage tool?

**29.** Propose an AI application from your own field where the cost of a false negative would clearly outweigh the cost of a false positive, or vice versa, and explain the practical implication.

**30.** If you had to advise a Nigerian institution considering a large, resource-intensive AI model, what questions from this chapter would you want them to answer first?

**31.** Using Section 15.13, explain how federated learning addresses the data sovereignty concerns raised in Chapter 16.

**32.** Describe what is shared between institutions in a federated learning setup, and, just as importantly, what is NOT shared.

**33.** Why might a model trained via federated learning across several institutions generalise better than a model trained on a single institution's data alone?

**34.** What practical or technical challenges might arise in coordinating a federated learning project across multiple Nigerian teaching hospitals?

**35.** How does federated learning relate to the subgroup generalisation concerns raised in Sections 11.5 and 11.7?

# 16 Data Ethics, Governance, and the Future of Health Data Science in Nigeria

## The Ethical Foundations of Health Data Research

Every technique in this book operates on data that ultimately originates from real people — patients, survey respondents, research participants — and the ethical obligations that attach to health research (respect for persons, beneficence, and justice, as classically formulated in the Belmont Report and operationalised in Nigeria through the National Code of Health Research Ethics and institutional Health Research Ethics Committee review) apply with undiminished force to secondary data analysis, machine learning, and big data projects, not only to primary data collection. Informed consent, confidentiality, and the principle of minimal necessary data use should shape a data science project's design from its earliest stage, not be treated as a compliance checkbox appended at the end.

## Privacy, De-identification, and the Limits of Anonymisation

Removing direct identifiers (name, address, phone number) from a dataset is necessary but not sufficient for genuine de-identification, because a combination of seemingly innocuous quasi-identifiers (age, sex, LGA of residence, rare diagnosis) can

be sufficient to re-identify an individual, particularly in a small facility or a rare-disease dataset — a risk this book's own worked examples (dermatoglyphic studies, small-area NTD surveillance) are directly exposed to. Techniques such as k-anonymity (ensuring every combination of quasi-identifiers is shared by at least k records), aggregation to a coarser geographic or age band, and, for the most sensitive projects, formal differential privacy methods, should be considered proportionate to the sensitivity of the data and the size of the population it describes.

## Algorithmic Bias and Fairness in Health Prediction

A predictive model (Chapter 11) trained on historical data inherits any bias present in how that data was generated — if historical access to a diagnostic test or a specialist referral was itself unequal across geography or socioeconomic status, a model trained on outcomes conditional on that access will reproduce and can even amplify the inequity if deployed uncritically. Auditing a health prediction model's performance across relevant subgroups (as introduced in Section 11.5) is not an optional extra step but a core ethical obligation whenever a model's output will influence who receives clinical attention, referral, or resources, and it should be conducted and reported alongside, not instead of, overall performance metrics.

## Data Sharing, Sovereignty, and Institutional Capacity

Nigerian health data increasingly flows into international research collaborations, global health databases, and commercial cloud infrastructure, raising questions of data sovereignty (who controls, and benefits from, data collected within Nigeria) that a working data scientist should be able to discuss with collaborators and institutional research offices: appropriate material/data transfer agreements, local capacity-building commitments (co-authorship, training, technology transfer) as a condition of collaboration, and, wherever feasible, retention of a locally accessible copy of any

dataset generated from Nigerian participants or facilities. Building durable in-country capacity for the full data science pipeline described in this book — not only data collection, but analysis, modelling, and interpretation — is itself a matter of research equity, not merely of technical training.

## Worked Example

### A K-Anonymity Check on a Small Dataset

Consider a de-identified dataset (names removed) of 40 leishmaniasis patients from a single LGA, with quasi-identifiers age band (0-9, 10-19, 20+), sex, and ward of residence. Cross-tabulating these three quasi-identifiers reveals that one combination — female, age 0-9, Ward C — has exactly one matching record. This single record fails even 2-anonymity (the weakest reasonable standard, requiring at least two records per combination) and is at meaningful risk of re-identification by anyone in the community who knows a young girl in Ward C was recently treated for leishmaniasis. The appropriate remedy is not simply to delete that one row (which would create a different problem — silently biasing the dataset) but to coarsen the offending quasi-identifiers, for example by widening the age band or aggregating ward of residence to a larger administrative unit, until every combination is shared by at least k = 5 records, a commonly used minimum threshold in small-area public health data releases.

*R example*

```
# R: checking k-anonymity across quasi-identifier combinations
library(dplyr)

k_anon_check <- patient_data %>%
  count(age_band, sex, ward, name = "n_matching_records") %>%
  arrange(n_matching_records)

cat("Smallest group size (k):", min(k_anon_check$n_matching_records), "\n")
print(k_anon_check %>% filter(n_matching_records < 5))   # combinations failing k=5
```

## Worked Example: A Subgroup Fairness Audit

Returning to the triage risk model introduced in Chapter 11: an overall AUC of 0.85 sounds strong, but a subgroup audit reveals AUC of 0.88 for patients presenting at tertiary hospitals (where the training data was concentrated) versus only 0.74 for patients presenting at primary health centres — and, more concerning still, a calibration check shows the model under-predicts risk by an average of 12 percentage points specifically for primary-health-centre patients. Deploying this model unmodified at primary health centres, where it was least well validated, would mean the facilities with typically the fewest resources for clinical judgement to compensate for a flawed score are also the ones receiving the least reliable predictions — precisely the equity concern raised in Section 16.3, made concrete with numbers rather than left as an abstract worry.

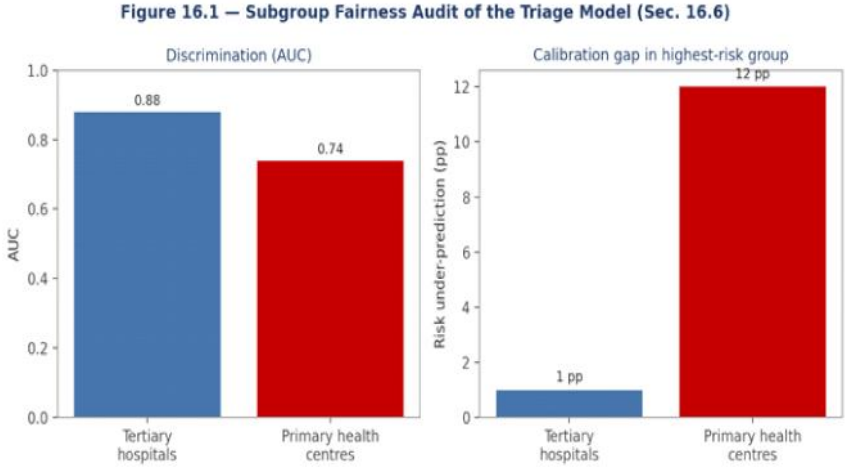



*Figure 16.1 — Subgroup fairness audit of the triage model (Section 16.6).*

*PYTHON example*

```
# Python: a subgroup AUC and calibration audit
from sklearn.metrics import roc_auc_score

for facility_type in ["tertiary", "primary"]:
    subset = triage_df[triage_df["facility_type"] == facility_type]
    auc = roc_auc_score(subset["severe_outcome"], subset["predicted_risk"])
    mean_predicted = subset["predicted_risk"].mean()
    mean_observed = subset["severe_outcome"].mean()
    print(f"{facility_type}: AUC={auc:.2f}, predicted={mean_predicted:.1%},
observed={mean_observed:.1%}")
```

## Looking Forward

The chapters of this book have moved from foundational probability and statistical inference, through classical and modern regression, survival analysis, time series forecasting, machine learning, unsupervised learning, mechanistic infectious disease modelling, spatial epidemiology, and big data/AI methods, consistently anchored in Nigerian health data and health system realities. The frontier of this field continues to move — toward richer integration of routine health system data with research cohorts, toward federated learning approaches that allow model training across institutions without centralising sensitive data, and toward AI-assisted diagnostic and decision-support tools appropriately validated for low-resource settings — but the foundational discipline this book has tried to instil (rigorous attention to where data comes from, what it can and cannot support, and who is accountable for how it is used) will remain the essential and durable skill beneath whatever specific technique or tool comes next.

## Community Engagement in Data Governance

Beyond formal ethical approval, meaningful community engagement — informing the communities from which health data is collected about how it will be used, and, where feasible,

involving community representatives in decisions about data sharing and secondary use — is increasingly recognised as good practice in health data governance, particularly for sensitive research involving stigmatised conditions or key populations (Chapter 7's respondent-driven sampling contexts). This is not merely an ethical nicety: community trust, once damaged by a perceived misuse of health data, can take years to rebuild and can compromise the willingness of an entire community to participate in future essential public health research and surveillance activities.

## Extended Case Study

### Well-Intentioned Data Sharing Request, Reconsidered

A genuinely difficult, realistic ethics scenario worth working through in detail: a well-regarded international research consortium requests access to a Nigerian institution's de-identified dermatoglyphic (fingerprint pattern) research dataset, comparing children with a specific congenital condition against unaffected controls, for inclusion in a larger, multi-country meta-analysis that could meaningfully advance understanding of the condition's early biomarkers. The request is well-intentioned, scientifically legitimate, and could benefit the broader research community, including future Nigerian patients who might benefit from earlier or more accurate diagnosis as a result. It is also, on the specific details of this dataset, higher-risk than most typical secondary data sharing requests: the condition in question is rare enough, and the study's recruitment sites specific enough, that even a carefully de-identified dataset carries real re-identification risk of the kind discussed in Section 16.2, and the original informed consent obtained from participants' families, some years earlier, did not explicitly anticipate international data sharing as a possible future use.

Working through this scenario using the framework of this chapter, several considerations must be weighed together rather

than in isolation: whether the original consent language is broad enough to reasonably cover this new use, or whether a fresh ethical review and, ideally, a renewed consent process would be required; whether additional data transformation (coarsening age bands, aggregating recruitment site information beyond what the international meta-analysis strictly needs) could reduce re-identification risk to an acceptable level without compromising the scientific value of the shared data; and whether the collaboration agreement includes the kind of local capacity-building and co-authorship provisions discussed in Section 16.4, ensuring the Nigerian research team and institution receive appropriate recognition and benefit rather than simply serving as a data source for an externally-led publication. The eventual resolution in a case of this kind — often involving referral back to the original institutional Health Research Ethics Committee for a specific ruling on whether the original consent covers this new use, and negotiation of an explicit data use and co-authorship agreement before any data transfer occurs — illustrates that responsible data governance in practice is rarely a simple yes/no gate, but a structured process of working through genuine, substantive tensions between scientific value, participant protection, and research equity.

## Distinguishing Data Ethics Review from Data Protection Compliance

It is worth explicitly distinguishing two related but distinct obligations that a Nigerian health data science project must satisfy: research ethics review (obtained from an institutional Health Research Ethics Committee, assessing whether the research is scientifically justified and ethically conducted with respect to participants) and data protection compliance (governed in Nigeria by the Nigeria Data Protection Act 2023 and overseen by the Nigeria Data Protection Commission, concerned specifically with the lawful processing, storage, and transfer of personal data

regardless of research purpose). A project can satisfy one without automatically satisfying the other — a well-designed study with full ethical approval could still process personal data in a way that falls short of data protection requirements (for example, transferring identifiable data internationally without an adequate legal basis) — and a mature institutional data governance process should treat these as two separate, both-necessary checks, rather than assuming ethical approval alone covers all relevant legal and regulatory obligations.

## Whistleblowing and Reporting Data Misuse

This chapter has focused primarily on how to design and conduct a data science project ethically from the outset; it is equally important to briefly address what a data scientist should do upon discovering that data has already been misused — for example, finding that a colleague has shared identifiable patient data outside an agreed data use agreement, or that a dataset's findings have been selectively reported to serve a predetermined political or programmatic narrative rather than the actual evidence. Most Nigerian research institutions and international collaborations maintain a formal research integrity or ethics reporting channel specifically for this purpose, and using it — rather than remaining silent out of loyalty to a colleague or discomfort with confrontation, or conversely making an unverified public accusation — is both an individual ethical obligation and, in the long run, a protection for the broader credibility of the data science profession and the communities whose data and trust are ultimately at stake; a data scientist who is aware of this reporting pathway before they ever need it is considerably better positioned to act appropriately and promptly if the need does arise.

## Data Retention and Secure Disposal Policies

Health data governance (Section 4.5, and this chapter's broader treatment) must address not only how data is collected, stored,

and shared, but how and when it is eventually disposed of. A written data retention policy — specifying how long identifiable data will be kept before being either securely destroyed or irreversibly de-identified, consistent with the original consent obtained from participants and any applicable institutional or regulatory requirements — protects both the research participants (whose data does not persist indefinitely beyond its original justified purpose) and the institution (which is not left holding an ever-growing, ever-riskier archive of identifiable sensitive data with no defined end point), and should be established at a project's outset alongside its ethical approval, not improvised years later when a dataset's continued retention is questioned.

## Auditing an Inherited Dataset for Undocumented Ethical Debt

A researcher who inherits a dataset originally collected by a previous team, sometimes years earlier, faces a specific challenge this chapter calls undocumented ethical debt: the original ethical approval, consent language, and any data use restrictions may be poorly documented, ambiguous about permitted secondary uses, or simply unavailable. Before conducting any new analysis on an inherited dataset, a responsible researcher should attempt to locate and review the original ethical approval and consent documentation, confirm whether the planned new analysis falls within its originally approved scope, and, where this cannot be confirmed with reasonable confidence, seek fresh ethical guidance rather than assuming that physical possession of a dataset implies unrestricted permission to analyse and publish from it.

## A General-Purpose k-Anonymity Checker Function

The k-anonymity check in Section 16.5 is written for one specific set of three quasi-identifiers; a general-purpose version that accepts any list of quasi-identifying columns can be reused across every future dataset a research team handles, rather than being

rewritten from scratch each time, directly supporting the personal code library habit recommended in Section 17.12.

***PYTHON example***

```
# Python: a reusable k-anonymity checker for any set of quasi-identifiers
def check_k_anonymity(df, quasi_identifiers, k=5):
    group_sizes = df.groupby(quasi_identifiers).size()
    failing = group_sizes[group_sizes < k]
    return {"min_group_size": int(group_sizes.min()), "n_failing_groups": len(failing),
            "passes_k_anonymity": len(failing) == 0, "failing_groups": failing}

result = check_k_anonymity(patient_df, ["age_band", "sex", "ward"], k=5)
if not result["passes_k_anonymity"]:
    print(f"{result['n_failing_groups']} group(s) fail k=5 anonymity — coarsen a quasi-
identifier")
```

## Chapter Summary

- Research ethics principles (respect for persons, beneficence, justice) apply fully to secondary analysis and machine learning, not only primary data collection.
- De-identification requires attention to quasi-identifiers and re-identification risk (k-anonymity), not just removal of direct identifiers.
- Predictive models can inherit and amplify historical inequities in access to care; subgroup performance auditing is an ethical obligation, not an optional extra.
- Data sovereignty, fair collaboration agreements, and local capacity-building are matters of research equity in international health data collaborations.

The durable skill of this book is disciplined attention to data provenance and accountability, which persists across whatever specific technique comes next

**Mini-Project**

**Mini-Project 16: A Privacy and Fairness Audit**

*Conduct both a k-anonymity check and a subgroup fairness audit on a dataset and/or model of your choice.*

**Step 1.** Select a de-identified dataset with at least three plausible quasi-identifiers (e.g., age band, sex, geographic unit).

**Step 2.** Compute the smallest group size across all quasi-identifier combinations (Section 16.5) and identify any combinations that fail a chosen k threshold.

**Step 3.** Propose and apply a specific coarsening strategy (e.g., wider age bands, larger geographic units) to bring all combinations up to your chosen k, and report the trade-off in information lost.

**Step 4.** If you have a predictive model (from a previous chapter's mini-project or a new one), compute its performance (AUC and calibration) separately for at least two meaningful subgroups.

**Step 5.** Write a one-page ethics and equity memo summarising both audits and your recommendations for safe release or deployment.

**Deliverable:** *A k-anonymity check with a remediation plan, a subgroup performance audit table, and a one-page ethics and equity memo.*

## End-of-Chapter Exercises

1. A dataset removes patient names but retains exact age, LGA, and a rare diagnosis. Explain the re-identification risk this creates and propose a mitigation.
2. Describe a plausible scenario in which a clinical risk-prediction model trained on Nigerian tertiary hospital data could disadvantage patients from primary health centres.
3. What provisions would you want to see in a data-sharing agreement before contributing a Nigerian patient cohort dataset to an international research collaboration?
4. Using the worked example in Section 16.5, explain why simply deleting the single at-risk record (rather than coarsening the quasi-identifiers) would itself introduce a new problem into the dataset.
5. In Section 16.6, explain why the combination of lower AUC AND under-prediction of risk at primary health centres is a more serious equity concern than either finding alone.

**6.** Using the R code in Section 16.5 as a guide, describe what would happen to the minimum group size (k) if 'ward' were replaced with the coarser variable 'LGA' in the grouping.

**7.** A funder asks you to justify, in two or three sentences, why 'the data has already had names removed' is not a sufficient answer to a privacy review question. Write that justification.

**8.** Using Section 16.8, explain why community trust, once damaged, could compromise future surveillance efforts beyond the immediate research project in question.

**9.** Propose one concrete way a research team could meaningfully engage a community before collecting sensitive health data from its members.

**10.** Why might community engagement be especially important for research involving a stigmatised condition or a key population, compared to a general household health survey?

**11.** In the Section 16.9 case study, explain why the fact that the original consent did not anticipate international data sharing matters, even though the new use is scientifically legitimate.

**12.** Propose one specific data transformation that could reduce re-identification risk for this dataset without eliminating its scientific value for the meta-analysis.

**13.** Why should the resolution of this kind of request generally involve the original institutional Health Research Ethics Committee, rather than being decided solely by the researchers holding the data?

**14.** What would you want included in a data use and co-authorship agreement before agreeing to share this dataset internationally?

**15.** Explain why 'the request is well-intentioned and scientifically valuable' is necessary but not sufficient

justification for approving a data sharing request like this one.

**16.** Using Section 16.10, explain why a study could have full research ethics approval but still fail to meet data protection compliance requirements.

**17.** Under the Nigeria Data Protection Act 2023 framework referenced in Section 16.10, why would international data transfer be treated as a distinct issue requiring its own legal basis?

**18.** Propose one practical step an institution could take to ensure both ethical approval and data protection compliance are checked, rather than assuming one covers the other.

**19.** Explain the difference in focus between a Health Research Ethics Committee's review and a Data Protection Commission's concerns, in your own words.

**20.** Why is it important for a data governance process to treat these as 'two separate, both-necessary checks' rather than a single combined approval step?

**21.** Using Section 16.11, explain why remaining silent about discovered data misuse is described as an ethical problem, not just a matter of personal discomfort.

**22.** Why is making an unverified public accusation also described as an inappropriate response, alongside remaining silent?

**23.** What is the recommended channel, according to Section 16.11, for reporting a genuine concern about data misuse?

**24.** Why is it useful to know about this reporting pathway 'before you ever need it,' rather than only learning about it after discovering a problem?

**25.** Propose a realistic (anonymised or hypothetical) scenario where you would feel obligated to use such a reporting channel, and explain why.

**26.** Summarise, in your own words, why this chapter treats data ethics review and formal data protection compliance as related but distinct processes (Section 16.10).

**27.** Explain the common thread linking the whistleblowing discussion (Section 16.11) and the reconsidered data-sharing request case study (Section 16.9).

**28.** Why does this chapter argue that a 'well-intentioned' data request can still raise legitimate ethical concerns?

**29.** Propose a scenario from your own institution where distinguishing ethics review from data protection compliance (Section 16.10) would matter in practice.

**30.** If you personally observed a colleague misusing sensitive health data, what would this chapter suggest as an appropriate first step?

**31.** Using Section 16.13, explain what is meant by 'undocumented ethical debt' in an inherited dataset.

**32.** What specific steps should a researcher take before analysing a dataset originally collected by a different team years earlier?

**33.** Why is physical possession of a dataset not, by itself, sufficient justification for analysing and publishing from it?

**34.** Describe a plausible scenario where a planned new analysis of an inherited dataset would fall outside its original ethical approval's scope.

**35.** How does the concept of undocumented ethical debt relate to the data dictionary and metadata documentation practices introduced in Chapter 4?

# 17 Capstone: Case Studies and Framework for Independent Practice

## Purpose of this Chapter

This chapter draws together the full pipeline introduced in Chapter 1 through a set of condensed case studies, each illustrating how several of the book's methods combine in a single applied Nigerian health data science project. It closes with a practical checklist that a reader can apply to a new project of their own, before the book's final chapter turns to a dedicated, hands-on treatment of Python for disease epidemiology.

## Case Study A

### Malaria Diagnostics Across Global Fund-Supported States

A comparative analysis of RDT, microscopy, and PCR-confirmed malaria testing across thirteen Global Fund-supported states draws on: Chapter 2's understanding of HMIS versus archive-verified data provenance; Chapter 4's cleaning workflow for facility-month records; Chapter 5's exploratory comparison of TPR across diagnostic modalities and states; Chapter 7's hypothesis testing (and multiple-comparison correction) for state-by-state differences; and Chapter 8's logistic regression modelling of confirmed positivity as a function of diagnostic method, season, and state. The key finding that HMIS-derived TPR runs consistently higher than archive-verified TPR across all thirteen states is a

direct, policy-relevant product of the provenance-aware approach this book has emphasised throughout, with implications for how confidently routine data alone can support commodity forecasting decisions.

## Case Study B

### Leishmaniasis Baseline Mapping and Trend Modelling

A national neglected tropical disease baseline analysis combines Chapter 10's Poisson GLM trend modelling (quantifying the average annual percentage change in reported cases) with Chapter 14's choropleth mapping and small-area smoothing (visualising the spatial distribution of case rates by state, and stabilising estimates for low-population LGAs) and an explicit attribution stamp on every output, reflecting this book's insistence (Chapter 4) that every analytic output should be traceable to its exact data source and generation date — essential when the resulting maps and trend estimates feed directly into an elimination programme's implementation planning and donor reporting.

## Case Study C

### Adolescent Health Cohort Analysis

A two-wave longitudinal analysis of an adolescent health cohort draws on: Chapter 4's handling of missingness across survey waves (distinguishing MAR attrition from MNAR attrition tied to worsening symptoms, and applying multiple imputation); Chapter 5's exploratory comparison of symptom severity by household food security and school attendance; and Chapter 8's logistic and ordinal regression modelling of depressive symptom severity as a function of individual and household covariates, with careful attention (Chapter 4) to the correlated-observation structure induced by repeated measurement of the same adolescents across waves.

## Case Study D

### Respondent-Driven Sampling of a Key Population

An RDS study of female sex workers draws directly on Chapter 7's specialised RDS estimators (comparing Volz-Heckathorn and successor-of-successor estimates for robustness), combined with Chapter 8's regression modelling of correlates of a health outcome of interest within the recruited sample, illustrating that even a technically sophisticated sampling design does not exempt the analyst from the confounding and model-diagnostic discipline introduced earlier in the book, nor from the privacy safeguards of Chapter 16 given the sensitivity of the population involved.

## Case Study E

### A Predictive Triage Model, From Data to Deployment Decision

A synthetic but realistic capstone example: a research team builds a severe-malaria risk model at triage, following the full arc of this book. Chapter 4 cleans and audits the raw triage register; Chapter 5's EDA reveals a right-skewed parasite density variable and a suspiciously high missingness rate for respiratory rate at one facility (traced, on investigation, to a broken pulse oximeter during a specific month); Chapter 8 and Chapter 11 together compare a LASSO logistic regression against a random forest, using facility-grouped cross-validation (Section 11.7) to avoid an optimistic performance estimate; Chapter 16's subgroup audit (Section 16.6) reveals the model under-performs at primary health centres; and the project concludes not with deployment, but with a recommendation to collect a modest amount of additional primary-health-centre training data before proceeding — a genuinely realistic, and appropriately cautious, capstone outcome that models the kind of judgement this book has tried to cultivate throughout, rather than a premature declaration of success.

## A Practical Checklist for an Independent Project

- Have I written down the precise question before touching any data (Chapter 1)?
- Do I know exactly which data source(s) I am using, their unit of observation, and their known limitations (Chapter 2)?
- Have I documented my cleaning decisions in a reproducible script, not a manually edited spreadsheet (Chapter 4)?
- Have I explored the data (Chapter 5) before committing to a specific formal model?
- Have I chosen a method appropriate to my data structure and my goal — inference versus prediction (Chapters 7-13)?
- If I used machine learning, have I validated out-of-sample with a split that respects any grouping/clustering in the data (Chapter 11)?
- If my data or results have any spatial or temporal structure, have I accounted for it rather than assuming independence (Chapters 10 and 14)?
- Have I checked my model's assumptions and, where relevant, its performance across policy-relevant subgroups (Chapters 8, 11, and 16)?
- Have I obtained the necessary ethical approval and handled identifiable data appropriately, including a k-anonymity check where relevant (Chapter 16)?
- Can a colleague re-run my entire analysis from raw data to final output and get the same answer (Chapter 6)?

## Common Reasons Independent Projects Stall

- Underestimating data access lead time (Section 2.9), leaving too little time for the actual analysis once approvals finally come through.

- Skipping the exploratory analysis stage (Chapter 5) and moving directly to a formal model, only to discover a basic data quality problem much later that requires restarting the analysis.
- Choosing an unnecessarily complex method (a deep learning model for a 300-patient dataset, per Chapter 15's caution) that consumes disproportionate time relative to its benefit over a simpler, well-validated alternative.
- Failing to set up a reproducible project structure (Chapter 6) from the start, leading to confusion later about which script or dataset version produced a given result.

## Extended Case Study

### When a Capstone Analysis Contradicted the Initial Hypothesis

A final, instructive capstone example, chosen specifically because its outcome contradicted what the research team initially expected: a student project set out to test the hypothesis that a state's community health worker programme, which had been rapidly scaled up over an 18-month period, was responsible for an observed improvement in early ANC (antenatal care) attendance rates in the areas where the programme was deployed. Following this book's full pipeline — careful data provenance work distinguishing programme-deployment LGAs from comparison LGAs (Chapter 2), thorough exploratory analysis (Chapter 5) revealing that programme-deployment LGAs already had somewhat higher baseline ANC attendance before the programme began (a warning sign for confounding by pre-existing area characteristics, per Chapter 8), and a difference-in-differences regression model explicitly designed to account for this pre-existing gap rather than a simple before-after or treatment-vs-comparison comparison alone — the fully adjusted analysis found no statistically significant additional improvement in early ANC

attendance attributable to the programme beyond what would have been expected from the areas' pre-existing trajectory and other confounding trends.

This finding was, initially, unwelcome to a research team that had expected (and to some extent hoped) to demonstrate the programme's effectiveness. Rather than searching for an alternative specification or subgroup that would produce the originally-expected positive result (a temptation this book has warned against throughout, particularly in Chapter 7's discussion of multiple comparisons and Section 7.10's discussion of honestly reported null results), the team instead reported the null finding plainly, together with a careful discussion of its limitations (a relatively short 18-month follow-up window, and the possibility that the programme's benefits, if any, might operate through pathways other than early ANC attendance specifically, such as improved attendance at later ANC visits or improved postnatal care uptake, which the available data could not directly test). This honestly-reported, methodologically rigorous null result was, in the assessment of the student's supervisory committee, ultimately a stronger piece of work than a less careful analysis that had uncritically confirmed the initially expected positive finding would have been — precisely because it demonstrated the full analytical discipline this book has argued for throughout, applied faithfully even when its conclusion was not the one the research team had hoped to reach.

## A Note on Working with Co-Authors and Supervisors

Nearly every substantial project following this book's pipeline will involve at least one co-author or supervisor, and the reproducibility and documentation practices emphasised throughout this book (Chapter 6 especially) pay a second dividend specifically in this collaborative context: a supervisor or co-author who can read a clearly commented, version-controlled script and

its accompanying data provenance notes can review and contribute to a project far more efficiently, and with far more genuine understanding of what was actually done, than one who receives only a final results table with no visibility into the underlying process. Establishing early in a collaboration who is responsible for which stage of the pipeline (Section 1.2), and agreeing on a shared, version-controlled repository structure from the outset, prevents the common and avoidable friction of two collaborators independently producing slightly different versions of 'the' cleaned dataset or 'the' final model, each unaware that the other has made different, undocumented choices somewhere along the way.

## Presenting Findings to a Non-Technical Audience

Every case study and mini-project in this book has emphasised technically rigorous analysis; the final, and in many ways most consequential, step of the pipeline introduced in Chapter 1 is communicating that analysis's findings to an audience — an LGA health officer, a State Ministry of Health technical committee, a community leader — who may have little to no formal statistical training but who holds the actual authority or influence to act on the findings. Effective communication in this setting means leading with the practical implication rather than the statistical method ('this LGA needs additional net distribution before the rainy season' rather than 'the Poisson regression coefficient for LGA was significant at $p<0.05$'), using a small number of carefully chosen visualisations rather than a dense slide of tables, being explicit and honest about the confidence level and limitations of a finding without using so much technical hedging language that the message becomes unclear, and being prepared to answer 'so what should we actually do about it' as the natural and expected follow-up question to any presented finding — a skill this book's case studies (Section 17.2 through 17.6) have modelled throughout, but which ultimately develops most through direct

practice presenting real findings to real, non-technical stakeholders.

## Building a Reusable Personal Code Library Across Projects

A natural progression for a reader who has completed several of this book's mini-projects is to notice recurring code patterns — a cleaning function, a reporting template, a Kaplan-Meier plotting helper — and consolidate them into a personal, reusable code package rather than re-writing similar code from scratch for each new project. Both R (via a lightweight personal package built with 'devtools'/'usethis') and Python (via a simple installable local package) support this pattern well, and maintaining such a library, with basic documentation and version control (Chapter 6), compounds in value over a research career: each new project both benefits from and contributes back to an increasingly capable, personally-tailored analytic toolkit, which is one of the most practically valuable habits an independent Nigerian health data scientist can build over time.

## Writing the Methods Section

### Translating Code into Prose

A final practical skill this book has implicitly built toward throughout its many worked examples is translating a completed analysis — code, output, and all — into the concise, precise prose of a methods section suitable for a thesis chapter or journal submission. A strong methods section states the data source and study design, describes cleaning and exclusion decisions with their rationale (Chapter 4), names the exact statistical or machine learning method used with enough detail for replication (Chapters 7-15), and states how missing data, multiple comparisons, or model validation were handled — essentially a prose summary of the same provenance-aware, transparent decisions this book has

asked readers to make and document throughout every mini-project, now compressed into the formal register expected of a scientific publication.

## A Script to Scaffold a New Reproducible Project

The checklist in Section 17.7 is easiest to follow when a project's directory structure enforces good habits from the very first commit, rather than being retrofitted later. A short scaffolding script that creates the standard folders recommended throughout Chapter 6 (raw data, cleaned data, scripts, outputs), an empty data dictionary template (Section 4.8), and an initial Git repository in a single command removes any excuse for skipping these steps under time pressure at the start of a new project.

***PYTHON example***

```
# Python: a one-command project scaffolding script
import os, subprocess

def scaffold_project(project_name):
    folders = ["data/raw", "data/clean", "scripts", "outputs", "docs"]
    for folder in folders:
        os.makedirs(os.path.join(project_name, folder), exist_ok=True)
    with open(os.path.join(project_name, "docs/data_dictionary.csv"), "w") as f:
        f.write("variable,definition,type,allowed_range,source\n")
    subprocess.run(["git", "init", project_name], check=True)
    print(f"Scaffolded reproducible project at ./{project_name}")

scaffold_project("malaria_tpr_reconciliation_2026")
```

## Chapter Summary

- Real Nigerian health data science projects combine methods from across this book's chapters, not a single technique in isolation.
- Provenance-aware analysis (Chapter 2's core lesson) recurs as the key to interpreting each case study correctly.
- A capstone project's most valuable outcome is sometimes an appropriately cautious recommendation not to deploy yet, rather than a premature declaration of success.

- A short, disciplined checklist — question, data, cleaning, exploration, method choice, validation, structure, assumptions, ethics, reproducibility — is a practical summary of the entire book.

**Mini-Project**

**Mini-Project 17: A Full Capstone Project Proposal**

*Design (and, as far as feasible, execute) a complete capstone project of your own, applying the checklist in Section 17.7 end to end.*

**Step 1.** Choose a Nigerian health data question of genuine interest to you, distinct from the five case studies in this chapter.

**Step 2.** Draft a one-page project proposal identifying your data source(s), cleaning plan, exploratory questions, and intended formal analysis method(s), citing the relevant chapters.

**Step 3.** Execute as much of the pipeline as your data and time allow, from cleaning through to a formal analysis or model.

**Step 4.** Apply the ten-point checklist (Section 17.7) explicitly to your own project, noting any items you could not fully satisfy and why.

**Step 5.** Conclude with a short, honest statement of your project's current limitations and what would be needed before its findings could inform a real decision.

**Deliverable:** *A one-page project proposal, whatever analysis you were able to complete, and a completed checklist with honest self-assessment.*

## End-of-Chapter Exercises

1. Choose a health data project from your own work. Apply the ten-point checklist in Section 17.7 and identify the two items you are currently weakest on.
2. For Case Study A, propose one additional analysis (drawing on any chapter of this book) that would strengthen the policy conclusions.
3. Design, in outline form only, a new capstone case study from a Nigerian health data source not covered in Sections 17.2-17.6, specifying which chapters' methods it would draw on.

4. In Case Study E (Section 17.6), explain why recommending 'collect more data before deployment' is a more scientifically honest outcome than deploying the model with a known subgroup performance gap.

5. For Case Study C, explain how you would decide whether the adolescent cohort's follow-up attrition is more consistent with MAR or MNAR (Chapter 4), and what analysis you would run to investigate this.

6. For Case Study D, explain why privacy safeguards (Chapter 16) are especially important for an RDS study of a key population, compared to a general population household survey.

7. Write a short paragraph, in your own words, on which single habit from this entire book you consider most important to carry into your own future work, and why.

8. Using Section 17.8, identify which of the four listed reasons projects stall you consider most common in your own experience or observation, and explain why.

9. Explain how skipping Chapter 5's exploratory analysis stage could lead to a costly restart later in a project.

10. Propose one practical habit (from anywhere in this book) that would directly address one of the four stall points listed in Section 17.8.

11. In the Section 17.9 case study, explain why the pre-existing difference in baseline ANC attendance between programme and comparison LGAs was a warning sign that needed to be addressed before drawing conclusions.

12. Why was a difference-in-differences approach more appropriate here than a simple before-after or treatment-vs-comparison comparison alone?

**13.** Explain why the research team's supervisory committee judged the honestly reported null result to be stronger work than an uncritical positive finding would have been.

**14.** Propose one additional outcome (beyond early ANC attendance) that this programme's benefits might plausibly show up in, and how you would test for it.

**15.** What is the risk, for the research team's own credibility, of searching for an alternative specification that would produce the originally hoped-for positive result?

**16.** Using Section 17.10, explain why a co-author who can read a well-documented, version-controlled script can contribute more effectively than one who only sees a final results table.

**17.** Propose one specific agreement a research team should make at the start of a collaboration to avoid the 'two different versions of the cleaned dataset' problem described in Section 17.10.

**18.** Why does establishing pipeline responsibilities early (Section 1.2) help prevent friction later in a collaborative project?

**19.** Describe a situation (real or plausible) where a lack of shared, version-controlled workflow caused confusion or wasted effort in a team project.

**20.** How would you raise this issue diplomatically with a co-author or supervisor who is used to working without version control?

**21.** Using Section 17.11, rewrite the example phrase 'the Poisson regression coefficient for LGA was significant at $p<0.05$' into a plain-language, action-oriented statement suitable for a non-technical LGA health officer.

**22.** Why is being prepared to answer 'so what should we do about it' described as a natural and expected follow-up question?

**23.** Explain the risk of using too much technical hedging language when communicating a finding's limitations to a non-technical audience.

**24.** Propose one visualisation choice (from Chapter 5) that would likely communicate a key finding more effectively to a community leader than a dense data table.

**25.** Why does this book argue that this communication skill develops mainly through direct practice, rather than through reading about it alone?

**26.** Summarise, in your own words, why this chapter's checklist (Section 17.7) is presented as a practical summary of the entire book rather than a new set of ideas.

**27.** Explain the common thread linking the contradicted-hypothesis case study (Section 17.9) and the non-technical presentation discussion (Section 17.11).

**28.** Why does this chapter emphasise working constructively with co-authors and supervisors (Section 17.10) as a data science skill in its own right?

**29.** Propose a hypothetical capstone project from your own field, and describe how you would present its key finding to a non-technical audience (Section 17.11).

**30.** If your own capstone analysis contradicted your initial hypothesis, as in Section 17.9's case study, how would you decide what to report?

**31.** Using Section 17.13, list the four key elements a strong methods section should include, according to this chapter.

**32.** Why is a methods section described as a ‘prose summary’ of the same decisions documented throughout this book’s mini-projects?

**33.** Choose one of your own mini-project write-ups from this book and draft a two-to-three sentence methods section based on it.

**34.** Why is stating how missing data and multiple comparisons were handled considered essential for a complete methods section?

**35.** What is the risk of omitting the rationale for a cleaning or exclusion decision from a published methods section?

# Part 6

## Python for Disease Epidemiology

# 18 Introduction to Python for Disease Epidemiology

## Why a Dedicated Python Chapter, and Why Last

Every preceding chapter has used both R and Python where each language's strengths were most relevant. This closing chapter takes the opposite approach: a single, sustained, standalone walkthrough of an epidemiological data analysis workflow entirely in Python, from a raw outbreak line list to a mapped, modelled, and forecast set of results. It is placed last deliberately, so that a reader who has worked through the whole book can now see nearly every concept — data cleaning, EDA, hypothesis testing, regression, time series, spatial analysis, and mechanistic modelling — reappear as a connected, idiomatic Python pipeline, using the specific libraries (pandas, numpy, matplotlib, scipy, statsmodels, lifelines, geopandas, scikit-learn) that make up the modern Python epidemiology toolkit. A reader who works primarily in R can use this chapter as a self-contained bridge into Python; a reader who has followed the Python examples throughout the book will find this chapter consolidates and extends them into a single coherent analysis.

## Setting Up

### The Python Epidemiology Toolkit

A minimal but sufficient Python environment for disease epidemiology includes: pandas and numpy for data manipulation;

matplotlib and seaborn for visualisation; scipy and statsmodels for statistical inference and regression; lifelines for survival analysis; scikit-learn for machine learning; geopandas for spatial data; and epyestim or a simple custom implementation for epidemic parameter estimation. All of these are open-source, free to use, and installable via pip, making Python a genuinely low-cost, high-capability option for a Nigerian research group or state disease surveillance unit without a budget for proprietary statistical software.

***PYTHON example***

```
# Terminal (not Python code): installing the core Python epidemiology toolkit
# pip install pandas numpy matplotlib seaborn scipy statsmodels lifelines scikit-learn
geopandas

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

print("pandas:", pd.__version__)
print("numpy:", np.__version__)
```

## Working with an Outbreak Line List

A line list — one row per case, with columns for identifying and clinical information — is the fundamental data structure of field epidemiology, whether for a Lassa fever outbreak investigation coordinated through SORMAS or a localised cholera cluster. The first task with any line list is always the same set of checks introduced in Chapter 4: dimensions, types, duplicates, and date plausibility (a symptom onset date after a reported death date is an internal-consistency error that must be resolved before any epidemic curve is built).

***PYTHON example***

```
# Python: loading and validating an outbreak line list
import pandas as pd

line_list = pd.read_csv("lassa_outbreak_linelist.csv", parse_dates=["onset_date",
"report_date", "outcome_date"])

print(line_list.shape)
print(line_list.dtypes)

# Internal consistency: onset date should never be after outcome date
bad_dates = line_list[line_list["onset_date"] > line_list["outcome_date"]]
print(f"Records with onset after outcome: {len(bad_dates)}")

# Duplicate case check
duplicates = line_list.duplicated(subset=["patient_id"]).sum()
print(f"Duplicate patient IDs: {duplicates}")
```

## Building and Interpreting an Epidemic Curve

The epidemic curve (epicurve) — a histogram of case counts by date of symptom onset — is the single most important visualisation in an outbreak investigation, because its shape provides immediate clues about the likely mode of transmission: a single sharp peak suggests a common-source exposure (such as a contaminated water source), while a series of successive peaks spaced roughly one incubation period apart suggests ongoing person-to-person transmission. Python's pandas and matplotlib make constructing an epicurve, and overlaying key events (an intervention start date, for example), straightforward.

***PYTHON example***

```
# Python: building an epidemic curve with an intervention marker
import matplotlib.pyplot as plt

daily_counts = line_list.groupby(line_list["onset_date"].dt.date).size()

fig, ax = plt.subplots(figsize=(10, 5))
daily_counts.plot(kind="bar", ax=ax, color="firebrick", width=0.9)
ax.axvline(x=pd.Timestamp("2024-06-15"), color="steelblue", linestyle="--",
           label="Case isolation protocol introduced")
ax.set_xlabel("Date of symptom onset")
ax.set_ylabel("Number of cases")
ax.set_title("Epidemic Curve: Suspected Lassa Fever Cluster")
ax.legend()
plt.tight_layout()
plt.savefig("epicurve.png", dpi=200)
```

## Attack Rates, Case Fatality Ratios, and Their Confidence Intervals

Two of the most frequently reported (and frequently miscalculated) outbreak statistics are the attack rate (the proportion of an at-risk population that becomes a case) and the case fatality ratio, or CFR (the proportion of cases that die). Both are simple proportions, and both should always be reported with a confidence interval and, for the CFR in particular, with explicit attention to outcome-reporting delay: cases still hospitalised or under follow-up at the time of analysis have an unknown outcome, and including them in the denominator of a naive CFR calculation (deaths divided by all cases) will underestimate the true CFR, sometimes substantially, during an active outbreak.

***PYTHON example***

```
# Python: calculating attack rate and CFR with confidence intervals,
# correctly excluding cases with an unresolved (ongoing) outcome
from statsmodels.stats.proportion import proportion_confint

# Attack rate
population_at_risk = 45000
total_cases = len(line_list)
attack_rate = total_cases / population_at_risk
ar_ci = proportion_confint(total_cases, population_at_risk, method="wilson")
print(f"Attack rate: {attack_rate:.4%} (95% CI {ar_ci[0]:.4%} - {ar_ci[1]:.4%})")

# CFR, correctly restricted to cases with a resolved outcome
resolved = line_list[line_list["outcome"].isin(["died", "recovered"])]
deaths = (resolved["outcome"] == "died").sum()
naive_cfr = (line_list["outcome"] == "died").sum() / total_cases
resolved_cfr = deaths / len(resolved)
cfr_ci = proportion_confint(deaths, len(resolved), method="wilson")

print(f"Naive CFR (includes unresolved cases in denominator): {naive_cfr:.1%}")
print(f"Corrected CFR (resolved cases only): {resolved_cfr:.1%} (95% CI {cfr_ci[0]:.1%} -
{cfr_ci[1]:.1%})")
```

## Estimating the Effective Reproduction Number from Case Data

Chapter 13 introduced R0, the basic reproduction number, from a fitted mechanistic model. During an actual, ongoing outbreak,

epidemiologists more often need the effective reproduction number, Rt — the time-varying average number of secondary cases per infectious case, which can be estimated directly from an epidemic curve and an assumed serial interval distribution (the time between symptom onset in a case and their infector), without needing to fit a full SEIR model. A simple, widely used approach (a discretised version of the Cori et al. method) estimates Rt for each time window as the ratio of new cases in that window to the sum of previous cases weighted by the serial interval distribution.

***PYTHON example***

```
# Python: a simplified Cori-method style Rt estimate
import numpy as np
from scipy.stats import gamma

daily_cases = daily_counts.values
serial_interval_mean, serial_interval_sd = 4.5, 2.0   # example serial interval, in days

# Discretised serial interval distribution (gamma approximation)
shape = (serial_interval_mean / serial_interval_sd) ** 2
scale = (serial_interval_sd ** 2) / serial_interval_mean
si_distribution = gamma.pdf(np.arange(1, 15), a=shape, scale=scale)
si_distribution /= si_distribution.sum()

def estimate_rt(cases, si_dist, window=7):
    rt_estimates = []
    for t in range(len(si_dist), len(cases)):
        infectiousness = sum(cases[t - k] * si_dist[k - 1] for k in range(1, len(si_dist) +
1))
        recent_cases = cases[max(0, t - window + 1): t + 1].sum()
        rt_estimates.append(recent_cases / (infectiousness * window) if infectiousness > 0
else np.nan)
    return rt_estimates

rt_series = estimate_rt(daily_cases, si_distribution)
print("Most recent Rt estimate:", round(rt_series[-1], 2) if rt_series else "insufficient
data")
```

## Spatial Mapping of Cases in Python

Chapter 14 introduced spatial disease mapping conceptually and largely in R; the same choropleth and point-mapping tasks are equally well supported in Python via geopandas, which extends pandas with geometric operations and integrates directly with

matplotlib for static maps or with folium for interactive web maps suitable for sharing with a State Ministry of Health team without specialised GIS software.

*PYTHON example*

```
# Python: choropleth mapping of case rates by LGA with geopandas
import geopandas as gpd
import matplotlib.pyplot as plt

lga_boundaries = gpd.read_file("lga_boundaries.shp")
case_summary = line_list.groupby("lga").size().reset_index(name="cases")

merged = lga_boundaries.merge(case_summary, left_on="LGA_NAME", right_on="lga", how="left")
merged["cases"] = merged["cases"].fillna(0)
merged["rate_per_100k"] = (merged["cases"] / merged["population"]) * 100000

fig, ax = plt.subplots(figsize=(8, 8))
merged.plot(column="rate_per_100k", cmap="OrRd", linewidth=0.3, edgecolor="grey",
            legend=True, ax=ax)
ax.set_title("Suspected Lassa Fever Cases per 100,000 Population by LGA")
ax.axis("off")
plt.savefig("outbreak_map.png", dpi=200, bbox_inches="tight")
```

## A Compact SEIR Implementation for Field Use

Building on Chapter 13's SEIR model, this section packages the model into a single reusable Python function that a field epidemiology team can quickly re-run with updated parameters as an outbreak evolves, rather than re-deriving the model from scratch each time — exactly the kind of function-based, reproducible practice argued for in Chapter 6, now applied to a mechanistic rather than a statistical model.

***PYTHON example***

```
# Python: a reusable, field-ready SEIR function with basic outputs
import numpy as np
from scipy.integrate import odeint

def run_seir(N, I0, beta, sigma, gamma, days=180):
    def seir(y, t):
        S, E, I, R = y
        dS = -beta * S * I / N
        dE = beta * S * I / N - sigma * E
        dI = sigma * E - gamma * I
        dR = gamma * I
        return [dS, dE, dI, dR]

    y0 = [N - I0, 0, I0, 0]
    t = np.linspace(0, days, days + 1)
    result = odeint(seir, y0, t)
    peak_day = int(np.argmax(result[:, 2]))
    return {
        "R0": round(beta / gamma, 2),
        "peak_infectious": round(result[:, 2].max()),
        "peak_day": peak_day,
        "trajectory": result
    }

baseline = run_seir(N=200000, I0=5, beta=0.32, sigma=1/6, gamma=1/12)
print(f"R0 = {baseline['R0']}, peak infectious = {baseline['peak_infectious']} on day
{baseline['peak_day']}")
```

## A Simple Risk Prediction Model for Case Severity

Consolidating Chapters 8 and 11's regression and machine learning material into the outbreak context, a logistic regression (or, with more data, a random forest per Chapter 11) can predict which confirmed cases are at elevated risk of a severe outcome based on age, time from onset to presentation, and presenting symptoms — supporting the same kind of triage-prioritisation use case discussed throughout this book, now demonstrated end-to-end on outbreak line-list data specifically.

***PYTHON example***

```
# Python: a simple severity risk model fitted directly on outbreak line-list data
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score

features = line_list[["age", "days_onset_to_presentation", "bleeding_symptom",
                      "altered_consciousness"]].fillna(0)
outcome = (line_list["outcome"] == "died").astype(int)

X_train, X_test, y_train, y_test = train_test_split(features, outcome, test_size=0.25,
                                                    random_state=42, stratify=outcome)
model = LogisticRegression(max_iter=1000).fit(X_train, y_train)
predicted_risk = model.predict_proba(X_test)[:, 1]

print(f"Test AUC: {roc_auc_score(y_test, predicted_risk):.2f}")
print(dict(zip(features.columns, model.coef_[0].round(2))))
```

## Worked Example

### A Complete Mini-Outbreak Analysis, Start to Finish

To consolidate this chapter, consider a compact, fully worked scenario: a 90-day suspected cholera cluster in a riverine LGA. The complete Python workflow proceeds through every stage covered above: (1) load and validate the 340-record line list, finding and correcting 6 duplicate patient IDs and 2 records with implausible onset-after-outcome dates; (2) build the epicurve, revealing a single sharp peak around day 18 followed by a long, slowly declining tail — a pattern more consistent with an initial common-source exposure (a contaminated water point) followed by limited secondary person-to-person spread, rather than a sustained community-transmission pattern; (3) calculate an attack rate of 1.3% in the exposed population and a resolved-case CFR of 2.1% (95% CI 0.9%-4.3%); (4) estimate Rt, finding it falls below 1 by day 25, consistent with the declining tail observed in the epicurve and with the water-source remediation intervention implemented on day 20; (5) map cases by ward, finding via Moran's I (Chapter 14 logic, implemented in Python per Section 18.7's

tools) that cases are significantly spatially clustered around the two wards nearest the implicated water point; and (6) conclude, in a two-paragraph summary for the State Ministry of Health, that the evidence is most consistent with a point-source outbreak successfully contained by the day-20 intervention, while recommending continued surveillance for 4 further weeks (roughly two serial intervals) before formally declaring the cluster closed.

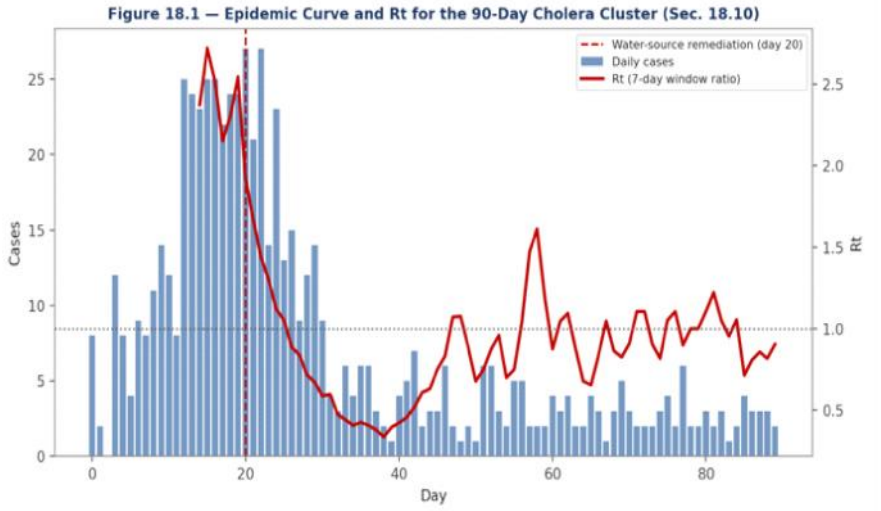



*Figure 18.1 — Epidemic curve and Rt for the 90-day cholera cluster (Section 18.10).*

## Deploying a Python Analysis as a Shareable Tool

Once an analysis pipeline of the kind built in this chapter has proven useful, converting it from a one-off script into a lightweight, shareable tool substantially increases its value to a broader team. Simple options include wrapping the core functions into a command-line script that a non-programmer field team member can run with a single command and a new data file, or building a minimal interactive dashboard (using a library such as Streamlit) that displays the epicurve, attack rate, and Rt estimate for

whichever line list file is uploaded — turning what was originally a personal analysis script into institutional infrastructure that survives beyond the original analyst's involvement in the project, directly supporting the kind of durable, in-country data science capacity argued for in Chapter 16.

## Extended Case Study

### A Python Pipeline That Caught Its Own Data Entry Error

A fitting closing case study for this book's final chapter, illustrating how the validation habits introduced in Chapter 4 and reinforced throughout this Python chapter can catch errors that would otherwise propagate silently into a final report: during the Section 18.10-style analysis of a real (anonymised for this account) suspected cholera cluster, the line-list validation step (Section 18.3) flagged three records where the recorded onset date fell after the recorded outcome (death) date — an internal impossibility. Rather than simply dropping these three records (a defensible but information-discarding response), the field team was asked to check the original paper case investigation forms, and discovered that all three were simple date-transcription errors: the day and month had been transposed when the data was entered into the electronic line list (for example, an intended 3 June being entered as 6 March), a common and specific error pattern in contexts where both DD/MM/YYYY and MM/DD/YYYY date formats circulate in different training materials and software defaults.

Correcting these three transcription errors (rather than excluding the records) changed the shape of the epidemic curve's early days meaningfully: two of the three corrected records fell within the first week of the outbreak, meaning the true earliest known case was four days earlier than the uncorrected data had suggested, which in turn shifted the team's estimate of the outbreak's likely true index exposure window earlier as well, with

direct implications for which water sources and community events fell within the revised, plausible exposure period under investigation. This case study exemplifies, in miniature, the argument threaded through this entire book: a validation check that might seem like a minor, mechanical data-cleaning formality (Section 18.3's date consistency check) can, when taken seriously and followed up rather than silently worked around, materially change a field epidemiological investigation's substantive conclusions — precisely the kind of provenance-aware, undismissive attention to data quality that Chapter 1 argued, at the very start of this book, was the single most important habit a health data scientist can bring to their work.

## A Brief Comparison

### When Python's Advantages Are Most Decisive

Having now worked through a complete Python epidemiology pipeline, it is worth stepping back and stating plainly where Python's particular advantages, relative to R, are most decisive for a Nigerian field epidemiology or health data science team, complementing the more even-handed comparison in Section 6.1. Python's advantage is most decisive when a pipeline must eventually be integrated into a broader software system (a hospital information system, a mobile data collection app's backend, an automated dashboard or API), because Python's general-purpose programming ecosystem and web-framework support are considerably more mature than R's for this kind of production software engineering task; and when a project's core analytic need shifts toward large-scale machine learning, deep learning (Chapter 15), or geospatial web-mapping (Section 18.7's folium-based interactive maps), where Python's library ecosystem is generally deeper and more actively maintained. Where a project's core need remains classical statistical inference, survival analysis, or complex survey design (Chapters 7 and 9 especially), R's longer-

established statistical ecosystem often remains the more natural first choice — reinforcing, rather than contradicting, this book's consistent position that the right tool depends on the specific task at hand, not on a fixed allegiance to one language.

## A Final Reflection

### Closing the Loop Back to Chapter 1

This book opened, in Chapter 1, with the argument that data science for the health sciences is best understood as a connected pipeline rather than a collection of disconnected techniques, and closes, in this final Python chapter, having walked through that entire pipeline concretely, in a single programming language, on a single realistic outbreak scenario. The Python tools introduced in this chapter — pandas for data manipulation, matplotlib for visualisation, scipy and statsmodels for inference, geopandas for spatial analysis, scikit-learn for prediction — are not fundamentally different in their underlying statistical and epidemiological logic from the R tools used throughout the rest of this book; they are simply a different, equally capable set of instruments for applying the same disciplined thinking this book has argued for throughout.

The most durable outcome a reader can take from this book is not fluency in any specific function call or package, both of which will inevitably change and be superseded over the course of a career, but the underlying habits of mind this book has tried to model consistently across all eighteen chapters and every worked example: ask where the data came from before trusting a number derived from it; distinguish a statistically significant result from a practically important one; validate a model on data it has not seen, respecting whatever natural structure the data actually has; and communicate findings, and their genuine limitations, honestly to the people who will act on them. A reader who internalises these habits will be well equipped to apply whatever specific tool

— R, Python, or a language not yet invented — the next stage of Nigerian health data science comes to rely on.

## Where to Go Next

### Advanced Resources for Python in Epidemiology

Readers who complete this chapter and want to extend their Python epidemiology skills further are encouraged to explore: the ‘epyestim‘ and ‘EpiEstim‘-equivalent Python packages for more rigorous Rt estimation than the simplified approach in Section 18.6; the ‘PyMC‘ and ‘Stan‘-via-‘cmdstanpy‘ ecosystems for full Bayesian mechanistic model fitting, extending the point-estimate parameter fitting of Section 18.8 to a full posterior distribution over parameters; and the ‘geemap‘/‘Google Earth Engine‘ Python API for large-scale environmental covariate extraction to complement Chapter 14’s spatial methods. None of these tools change the fundamental analytic discipline this book has tried to teach — provenance-aware, validated, honestly-communicated analysis — but each extends the technical sophistication available to a reader ready to move beyond this book’s introductory treatment.

## A Final Worked Example

### Comparing R and Python Outputs on the Same Dataset

As a closing demonstration, fitting the same logistic regression specification (Section 18.9’s severity risk model) in both R’s ‘glm()‘ and Python’s ‘scikit-learn‘/‘statsmodels‘ on the identical dataset should, and in a correctly implemented analysis will, produce numerically identical or near-identical coefficient estimates — a useful, concrete check that an analyst has correctly translated a model specification between the two languages, and a reassuring practical demonstration that the choice between R and Python, emphasised as a matter of ecosystem convenience throughout this book (Section 6.1), does not in itself change the underlying statistical answer a correctly specified model

produces. Any meaningful discrepancy between the two languages' output on the same model and data should prompt the analyst to check for a specification difference (a different reference category, a different default optimisation method) rather than assuming one language is simply 'wrong.'

## Packaging this Chapter's Functions into a Reusable Module

Every function introduced across this chapter — line-list validation, epicurve construction, attack rate and CFR calculation, Rt estimation — has so far lived in an ad hoc sequence of notebook cells. Collecting them into a single, importable Python module (a plain '.py' file with each function defined and documented) is the final, small step that turns this chapter's individually useful pieces into the kind of reusable personal code library recommended in Section 17.12, allowing a future outbreak investigation to start from 'import outbreak_toolkit as ot' rather than re-copying code from this book each time.

***PYTHON example***

```
# outbreak_toolkit.py — a minimal packaging of this chapter's core functions
"""Reusable field epidemiology functions developed in Chapter 18."""
import pandas as pd
from statsmodels.stats.proportion import proportion_confint

def validate_linelist(df):
    issues = []
    if (df["onset_date"] > df["outcome_date"]).any():
        issues.append("onset after outcome")
    if df["patient_id"].duplicated().any():
        issues.append("duplicate patient IDs")
    return issues

def attack_rate_with_ci(n_cases, population):
    rate = n_cases / population
    ci = proportion_confint(n_cases, population, method="wilson")
    return {"attack_rate": rate, "ci_low": ci[0], "ci_high": ci[1]}

def resolved_cfr(df):
    resolved = df[df["outcome"].isin(["died", "recovered"])]
    deaths = (resolved["outcome"] == "died").sum()
    return deaths / len(resolved) if len(resolved) else float("nan")

# Usage elsewhere: from outbreak_toolkit import validate_linelist, attack_rate_with_ci,
resolved_cfr
```

## Chapter Summary

- Python's open-source epidemiology toolkit (pandas, numpy, matplotlib, scipy, statsmodels, lifelines, scikit-learn, geopandas) supports a complete field epidemiology workflow at no licensing cost.
- A line list requires the same validation discipline (dimensions, types, duplicates, date plausibility) introduced in Chapter 4, applied specifically to outbreak case data.
- The shape of an epidemic curve provides immediate, visual evidence about likely transmission mode (point-source vs. person-to-person).
- Attack rate and case fatality ratio must be reported with confidence intervals, and CFR calculations during an active outbreak must correctly handle cases with unresolved outcomes.
- The effective reproduction number Rt can be estimated directly from case data and a serial interval distribution, without needing to fit a full mechanistic model.
- Spatial mapping, mechanistic modelling, and predictive risk modelling — introduced separately in earlier chapters — combine naturally into a single Python-based outbreak analysis workflow.
- A complete outbreak analysis concludes with a plain-language epidemiological interpretation (mode of transmission, intervention impact, ongoing risk) — not merely a set of disconnected statistics.

**Mini-Project**

**Mini-Project 18 (Capstone): A Complete Python Outbreak Investigation**

*Conduct a complete, end-to-end outbreak analysis in Python, following the structure of Section 18.10, on a simulated or real line-list dataset.*

**Step 1.** Simulate or obtain a line-list dataset of at least 200 cases with onset date, outcome, outcome date, age, geographic unit, and at least two clinical symptom variables.

**Step 2.** Validate the line list (Section 18.3), reporting any duplicates or date-consistency errors found and how you resolved them.

**Step 3.** Build and interpret an epidemic curve (Section 18.4), commenting on the likely mode of transmission suggested by its shape.

**Step 4.** Calculate the attack rate and a correctly-handled case fatality ratio with confidence intervals (Section 18.5).

**Step 5.** Estimate Rt over the course of the outbreak (Section 18.6) and relate its trajectory to any known or assumed intervention timing.

**Step 6.** Produce a spatial map of case rates by geographic unit (Section 18.7).

**Step 7.** Fit a simple severity risk model (Section 18.9) and report its cross-validated AUC.

**Step 8.** Write a two-paragraph epidemiological summary, in the style of Section 18.10, suitable for a State Ministry of Health situation report.

**Deliverable:** *A complete, runnable Python notebook or script covering all eight steps, plus the two-paragraph written summary.*

## End-of-Chapter Exercises

1. Explain, using the shape of an epidemic curve, how you would distinguish a point-source outbreak from one driven by ongoing person-to-person transmission.
2. A line list has 50 confirmed cases, of which 12 have died, 30 have recovered, and 8 are still hospitalised with an unknown outcome. Calculate both the naive CFR (using all 50 cases as the denominator) and the corrected CFR (using only resolved cases), and explain which is more appropriate to report while the outbreak is still active.

**3.** Explain, in your own words, the difference between R0 (Chapter 13) and Rt (Section 18.6), and why Rt is generally more useful for monitoring an outbreak in real time.

**4.** Using the worked example in Section 18.10, explain why finding that cases cluster significantly around two specific wards supports the point-source hypothesis rather than the person-to-person transmission hypothesis.

**5.** Using the code in Section 18.5, recompute the corrected CFR if 4 more of the 'unresolved' cases in the example were to later die, and comment on how much the CFR estimate could still change as an outbreak resolves.

**6.** In Section 18.9, explain why a logistic regression's coefficients being reported (rather than only its AUC) is useful for a field epidemiology team trying to understand which symptoms most elevate risk of a severe outcome.

**7.** Using Section 18.6's Rt approach, explain what a sustained Rt estimate above 1 for several consecutive time windows would suggest to an outbreak response team, and what action this might prompt.

**8.** Using Section 18.11, explain one advantage of converting a personal analysis script into a shareable command-line tool for a field surveillance team.

**9.** Why might an interactive dashboard (e.g., built with Streamlit) be more useful than a script for a non-programmer team member monitoring an ongoing outbreak?

**10.** What risk arises if institutional infrastructure (like the tool described in Section 18.11) depends entirely on one analyst's continued involvement, and how might a team mitigate this risk?

**11.** In the Section 18.12 case study, explain how a day/month transposition error could occur even when data entry staff are experienced and careful.

**12.** Why did correcting the three flagged records (rather than excluding them) matter for the outbreak's exposure-window investigation specifically?

**13.** Propose one systems-level change (e.g., to the data entry form or software) that could prevent this specific type of date error from occurring in future outbreaks.

**14.** Why is it significant that the validation check in Section 18.3, introduced early in this chapter, was what surfaced this error rather than a later, more complex analysis step?

**15.** Write a two-sentence closing reflection, in your own words, connecting this final case study back to the very first chapter's argument about data provenance.

**16.** Using Section 18.13, explain why Python's advantage is most decisive when a pipeline must be integrated into a broader software system.

**17.** For which specific chapters or tasks in this book does Section 18.13 suggest R's ecosystem remains the more natural first choice?

**18.** Explain why this book takes the position that 'the right tool depends on the task,' rather than recommending one language over the other in general.

**19.** Propose a Nigerian health data science task not explicitly mentioned in Section 18.13 where you would lean toward Python, and justify your choice.

**20.** Propose a Nigerian health data science task where you would lean toward R instead, and justify your choice, drawing on any relevant earlier chapter.

**21.** Using Section 18.14, explain in your own words why this book argues fluency in a specific tool is less durable than the underlying habits of mind it has tried to teach.

**22.** Which of the four habits listed in Section 18.14 (provenance, significance vs. importance, out-of-sample validation, honest communication) do you personally find most challenging to apply consistently, and why?

**23.** Why does the chapter emphasise that R and Python are 'not fundamentally different in their underlying statistical and epidemiological logic,' despite their different syntax?

**24.** Explain how the final chapter's closing argument connects back specifically to the four-question checklist introduced in Section 1.9.

**25.** Write a short, personal closing reflection (3-5 sentences) on which single idea from this entire book you intend to apply first in your own work, and why.

**26.** Summarise, in your own words, why this chapter closes the book by returning explicitly to Chapter 1's central argument (Section 18.14).

**27.** Explain the common thread linking the self-correcting pipeline case study (Section 18.12) and the discussion of when Python's advantages are most decisive (Section 18.13).

**28.** Why does this chapter argue that a well-built validation check (as in Section 18.12) is as valuable as the analysis itself?

**29.** Propose one scenario from your own field where Python's specific strengths (Section 18.13) would make it clearly preferable to R, and one where the reverse would be true.

**30.** Having completed this book, what single practice do you now consider most essential to carry into your own future outbreak or health data analysis work?

**31.** Using Section 18.16, explain why fitting the same model specification in R and Python on identical data should produce numerically similar results.

**32.** If R and Python produced meaningfully different coefficients for the same model and data, what would you check first?

**33.** Why is this cross-language check a useful practical habit for an analyst who regularly works in both R and Python?

**34.** Name one common source of a legitimate (not erroneous) discrepancy between two languages' implementations of 'the same' model.

**35.** Reflecting on the whole of Chapter 18, what would you say is the single most transferable skill from this chapter to a context outside outbreak investigation specifically?

# Glossary of Key Terms

**Bayes' theorem**

A formula relating conditional probabilities, used here to link diagnostic test sensitivity/specificity to positive predictive value.

**Censoring**

In survival analysis, the condition where the exact event time is unknown because the event had not occurred (or was not observed) by the end of follow-up.

**Choropleth map**

A map that shades administrative units by the value of an indicator, such as disease case rate.

**Confounding**

A distortion of an exposure-outcome association caused by a third variable associated with both.

**DHIS2**

The District Health Information Software platform underlying Nigeria's National Health Management Information System (NHMIS).

**Design effect**

The factor by which a complex survey design (clustering, stratification) inflates variance relative to simple random sampling.

**Hazard ratio**
In survival analysis, the multiplicative effect of a covariate on the instantaneous risk of an event, estimated by a Cox model.

**MAR / MCAR / MNAR**
Missing at random / missing completely at random / missing not at random — mechanisms describing why data are missing.

**Odds ratio**
The exponentiated coefficient from a logistic regression, interpreted as a multiplicative change in the odds of an outcome.

**Overdispersion**
A situation where a count outcome's variance exceeds its mean, violating the Poisson assumption and motivating negative binomial regression.

**R0 (basic reproduction number)**
The expected number of secondary infections caused by one infectious individual in a fully susceptible population.

**Respondent-driven sampling (RDS)**
A peer-referral chain sampling method for populations without a sampling frame, requiring specialised network-based estimators.

**SEIR model**
A compartmental infectious disease model dividing a population into Susceptible, Exposed, Infectious, and Recovered states.

**Test positivity rate (TPR)**
The proportion of diagnostic tests performed that return a positive result.

**Attack rate**
The proportion of an at-risk population that becomes a case during a defined outbreak period.

**Basic reproduction number (R0)**

The expected number of secondary infections produced by one infectious individual in an otherwise fully susceptible population.

**Case fatality ratio (CFR)**

The proportion of confirmed cases that die from a given disease, ideally calculated using only cases with a resolved outcome.

**Cohen's d**

A standardised effect size expressing the difference between two group means in units of pooled standard deviation.

**Cross-validation**

A model validation technique that repeatedly trains and tests a model on different partitions (folds) of the data to estimate out-of-sample performance.

**Data dictionary**

A structured document defining every variable in a dataset, including its meaning, type, allowed range, and source.

**DBSCAN**

A density-based clustering algorithm that groups closely-packed points and labels sparse, isolated points as noise rather than forcing them into a cluster.

**Effective reproduction number (Rt)**

The time-varying average number of secondary cases produced per infectious case at a given point during an ongoing outbreak.

**Empirical Bayes smoothing**

A statistical technique that shrinks an unstable small-area estimate toward a more stable wider-area average, in proportion to its uncertainty.

**Epidemic curve (epicurve)**
A histogram of case counts by date of symptom onset, used to infer the likely mode of transmission in an outbreak.

**k-anonymity**
A privacy standard requiring that every combination of quasi-identifying variables in a dataset is shared by at least k records.

**LASSO regression**
A regularised regression method that penalises the size of coefficients and can shrink some coefficients exactly to zero, performing automatic variable selection.

**Moran's I**
A statistic testing whether the spatial distribution of a variable is more clustered than would be expected under complete spatial randomness.

**Multiple imputation**
A missing-data method that generates several plausible completed datasets, analyses each, and pools the results to reflect imputation uncertainty.

**Poisson regression**
A generalised linear model for count outcomes, assuming the outcome's variance equals its mean.

**Positive predictive value (PPV)**
The probability that a person who tests positive on a diagnostic test truly has the disease, which depends heavily on disease prevalence.

**Sensitivity and specificity**
Fixed properties of a diagnostic test: sensitivity is the probability of a positive result given true disease; specificity is the probability of a negative result given true absence of disease.

**Survival analysis**

The family of statistical methods for time-to-event data that correctly incorporates censored (incomplete) observations.

**Transfer learning**

A deep learning technique that re-uses a neural network already trained on a large general dataset, retraining only its final layers on a smaller, task-specific dataset.

**Change-point detection**

A statistical technique for identifying the point(s) in a time series at which its underlying mean, variance, or trend shifts abruptly.

**Federated learning**

A machine learning approach that trains a shared model across multiple institutions' datasets without centralising the underlying raw data.

**Nowcasting**

A correction applied to the most recent period of surveillance data to account for reporting delay, producing a more accurate estimate of the current true case count.

**Quantile regression**

A regression method that models a specified quantile (rather than the mean) of an outcome's conditional distribution.

**Landmark analysis**

A survival analysis technique that restricts analysis to a fixed time point, avoiding immortal time bias from covariates measured partway through follow-up.

# Selected References and Further Reading


National Malaria Elimination Programme, Federal Ministry of Health, Nigeria. National Guidelines for Diagnosis and Treatment of Malaria.

National Population Commission (NPC) [Nigeria] and ICF. Nigeria Demographic and Health Survey.

Nigeria Centre for Disease Control and Prevention. Integrated Disease Surveillance and Response (IDSR) Technical Guidelines.

Federal Ministry of Health, Nigeria. National Code of Health Research Ethics.

World Health Organization. Global Leishmaniasis Surveillance Reports.

Cox, D.R. Regression Models and Life-Tables. Journal of the Royal Statistical Society.

Kermack, W.O., McKendrick, A.G. A Contribution to the Mathematical Theory of Epidemics. Proceedings of the Royal Society A.

Rubin, D.B. Inference and Missing Data. Biometrika.

Volz, E., Heckathorn, D.D. Probability Based Estimation Theory for Respondent-Driven Sampling. Journal of Official Statistics.

Besag, J., York, J., Mollié, A. Bayesian Image Restoration, with Two Applications in Spatial Statistics. Annals of the Institute of Statistical Mathematics.

Anselin, L. Local Indicators of Spatial Association—LISA. Geographical Analysis.

Benjamini, Y., Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B.

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology.

Diggle, P.J. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. Chapman and Hall/CRC.

Ester, M., Kriegel, H.P., Sander, J., Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of KDD.

Fine, J.P., Gray, R.J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association.

Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Heckathorn, D.D. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. Social Problems.

James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning. Springer.

Lundberg, S.M., Lee, S.I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems.

National Bureau of Statistics and UNICEF. Multiple Indicator Cluster Survey (MICS), Nigeria.

Sweeney, L. k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B.

World Health Organization. Verbal Autopsy Standards: Ascertaining and Attributing Causes of Death.

Robins, J.M., Hernán, M.A., Brumback, B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology.

Efron, B., Tibshirani, R.J. An Introduction to the Bootstrap. Chapman and Hall/CRC.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. Bayesian Data Analysis. Chapman and Hall/CRC.

Rieke, N., et al. The Future of Digital Health with Federated Learning. npj Digital Medicine.

Diggle, P.J., Ribeiro, P.J. Model-Based Geostatistics. Springer.

# Appendix
# Answers to End-of-Chapter Exercise

This appendix provides model answers to every end-of-chapter exercise in this book, chapter by chapter. Answers are written to demonstrate the reasoning and, where a calculation is required, the arithmetic expected of a complete response; for open-ended exercises, the answer given illustrates the kind of response that would be considered complete rather than the single unique correct wording.

## Chapter 1

## Introduction to Data Science for the Health Sciences

***Q1.*** *List three sources of routinely collected health data available in your own state or institution. For each, identify one plausible way its data-generating process could bias a naive analysis.*

**A: Examples:** (a) DHIS2/NHMIS facility reports — bias risk: incomplete transcription of negative results inflates positivity; (b) NDHS/MICS survey data — bias risk: infrequent rounds mean estimates can be stale for a fast-moving indicator; (c) NCDC/SORMAS case-based surveillance — bias risk: reporting fatigue during a prolonged outbreak causes under-ascertainment in later weeks. Any three sources from Chapter 2's list are acceptable provided the bias mechanism is specific

to how that source generates data, not a generic 'data can be wrong' answer.

***Q2.*** *A colleague reports that 'malaria cases rose 40% this quarter' using only HMIS aggregate totals. Write three questions you would ask before accepting this claim.*

**A:** Reasonable questions include: (1) Is this comparing the same population/denominator across quarters, or could a change in testing volume alone explain the rise? (2) Has this indicator been validated against a source register recently, given the known tendency of HMIS-derived rates to run high? (3) Is the 40% increase driven by a genuine transmission increase, a reporting completeness change (e.g., more facilities reporting this quarter), or a seasonal pattern that would appear in past years' data too?

***Q3.*** *Sketch the data science pipeline (Section 1.2) for a question of your choice drawn from your own research area, naming a specific Nigerian data source at the collection stage.*

**A:** Answers will vary by chosen question, but a complete answer names a specific Nigerian data source (e.g., 'Global Fund diagnostic registers' or 'NDHS 2023') at the collection stage and shows at least one feedback loop from a later pipeline stage (e.g., exploratory analysis) back to an earlier one (e.g., collecting an additional variable).

***Q4.*** *Using the worked example in Section 1.5, calculate the percentage-point difference in TPR if the register audit had instead found 430 tests performed and 215 confirmed positive. Comment on whether the direction of the HMIS bias is the same as in the original example.*

**A:** New register audit: 430 tests, 215 confirmed positive, TPR = 215/430 = 50.0%. Original HMIS TPR was 61.8%, so the difference is 61.8% " 50.0% = 11.8 percentage points — larger than the original example's 7.6 points, but in the same direction

(HMIS higher than register-verified), consistent with the systematic under-transcription bias described in the chapter.

***PYTHON — supporting code***

```
# Recomputing the TPR difference under the revised register-audit figures
hmis_tpr = 210 / 340
audit_tpr = 215 / 430
diff_pp = (hmis_tpr - audit_tpr) * 100
print(f"HMIS TPR: {hmis_tpr:.1%}, Audit TPR: {audit_tpr:.1%}, Difference:
{diff_pp:.1f} pp")
# HMIS TPR: 61.8%, Audit TPR: 50.0%, Difference: 11.8 pp
```

***Q5.*** *Explain, in two or three sentences, why 'a large sample size' does not by itself protect an analysis from the kind of systematic bias illustrated in Section 1.5.*

**A:** A large sample size reduces random (sampling) error but does nothing to correct systematic (non-random) bias, such as the HMIS under-transcription pattern described in Section 1.5. In fact, a very large biased sample can produce a very narrow confidence interval around a wrong answer, giving false confidence — the problem is the data-generating process, not the sample size.

***Q6.*** *Run the R or Python code in Section 1.6 on a dataset of your choice (or a simulated one) and report the number of rows, columns, and any implausible TPR values found.*

**A:** Answers will vary depending on the dataset used, but should report the correct number of rows/columns from 'dim()'/ '.shape', correctly identify variable types, and correctly flag any TPR values outside [0,1] using the range/comparison checks shown in the code.

***PYTHON — supporting code***

```
# A minimal, complete run of the Section 1.6 inspection on a simulated dataset
import pandas as pd, numpy as np

np.random.seed(1)
sim = pd.DataFrame({
    "facility_id": [f"F{i%20:03d}" for i in range(300)],
    "tests_performed": np.random.poisson(40, 300),
    "confirmed_positive": np.random.binomial(40, 0.3, 300),
})
print(sim.shape)
sim["tpr"] = sim["confirmed_positive"] / sim["tests_performed"]
print("Implausible TPR rows:", int(((sim["tpr"] < 0) | (sim["tpr"] > 1)).sum()))
```

***Q7.*** *A donor agency asks you to justify why a full data provenance log (Mini-Project 1) is worth the time it takes to produce, given that 'the numbers are already in the spreadsheet.' Write a short, persuasive paragraph in response.*

**A:** A strong response notes that: (1) a provenance log costs a small, fixed amount of time up front but can prevent a much larger cost later (a retracted finding, a wrong policy recommendation); (2) 'the numbers are already there' says nothing about whether they were generated in a way that supports the comparison being made; and (3) documenting provenance is what allows a supervisor, reviewer, or future analyst to trust and reuse the analysis, which is itself valuable beyond the immediate task.

***PYTHON — supporting code***

```
def provenance_flags(df):
    flags = {}
    flags["duplicate_rows"] = int(df.duplicated().sum())
    flags["missing_facility_id"] = int(df["facility_id"].isna().sum())
    return flags

print(provenance_flags(sim))
# A concrete, minutes-long check like this is exactly the up-front cost a
provenance log buys back
```

***Q8.*** *Describe, in your own words, how the pipeline in Section 1.2 is better thought of as a loop than a straight line, using an example of when a later stage might send you back to an earlier one.*

**A:** The pipeline is a loop because an unexpected finding at a later stage (e.g., an implausible pattern discovered during exploratory analysis) very often requires returning to an earlier stage — for example, going back to collect an additional variable, or revisiting the original problem formulation because the available data cannot actually answer the question as first framed — rather than simply proceeding forward regardless.

***Q9.*** *Which of the three professional roles described in Section 1.3 most closely matches your own background, and which of the other two skill sets do you most need to strengthen?*

**A:** Answers will vary by individual background, but a complete answer names one of the three roles (biostatistician/ epidemiologist, M&E/health informatics professional, or computer scientist/data engineer) as the closest match and honestly identifies a specific skill gap relative to the other two (e.g., a biostatistician might need to strengthen programming/ software engineering practices; a computer scientist might need to strengthen epidemiological reasoning and research ethics).

***Q10.*** *Explain briefly why reading a chapter's worked example before attempting its exercises (Section 1.7) might be less effective for long-term learning than attempting the exercises first.*

**A:** Reading the worked example first can short-circuit the productive struggle of retrieving and applying the concept independently, which cognitive science research on learning ('the testing effect') suggests is what drives durable retention; simply recognising a worked solution when you see it feels like understanding but often does not transfer to being able to

produce the solution unaided later, which is the actual competency the exercises are meant to build.

***Q11.*** *In the Section 1.8 case study, explain in your own words why Analyst B's estimate, though less dramatic, was ultimately the more useful one for the Ministry of Health.*

**A:** Analyst B's estimate was more useful because it was defensible under scrutiny — a Ministry technical committee that later investigated the claim would find the adjusted 3-4 point estimate holds up, whereas the original 24% relative reduction claim would need to be walked back once the seasonal and reporting-completeness confounds were identified, damaging the programme's credibility more than a modest but accurate finding ever would.

***Q12.*** *Identify which specific step in Analyst B's investigation (reporting completeness, campaign timing, or seasonal decline) you think would be easiest to check first with the least additional data, and explain why.*

**A:** Checking reporting completeness is typically easiest to check first, since it only requires counting how many facilities submitted a report each month from data already at hand, whereas checking actual campaign roll-out dates by LGA or establishing the pre-existing seasonal pattern both require additional data gathering or a longer historical baseline.

***Q13.*** *If reporting completeness had instead stayed constant across the pre/post period in Section 1.8, would you expect the seasonal-decline confound to fully explain the observed fall on its own? Explain your reasoning.*

**A:** Even with constant reporting completeness, the seasonal decline confound alone would likely still explain a meaningful portion (roughly half, per the case study) of the observed fall, since the pre-period specifically spans the tail end of the rainy season when transmission naturally declines regardless of any

intervention — so the true intervention effect would still need to be estimated using a method that adjusts for season, not simply assumed away.

***Q14.*** *Describe one way a health programme could design its own monitoring so that Analyst A's kind of naive before-after comparison is less likely to be misleading in the first place.*

**A:** A programme could pre-register its evaluation plan (specifying in advance which comparison period, adjustment variables, and analysis method will be used) before launching an intervention, and could deliberately choose a comparison period or design (such as a stepped-wedge rollout) that helps distinguish intervention effects from seasonal and reporting-completeness trends from the outset.

***Q15.*** *Write a two-sentence explanation, suitable for a non-technical programme manager, of why 'the numbers went down after we started the programme' is not by itself proof the programme caused the improvement.*

**A:** A reasonable response: 'A drop in the numbers after we started the programme could be caused by many things besides the programme itself, like the usual seasonal pattern or more complete reporting — we need to check for those before we can say the programme caused the improvement.'

***Q16.*** *Using the four-question checklist in Section 1.9, apply it briefly to a dataset you have worked with recently, writing one short answer to each question.*

**A:** Answers will vary by chosen dataset, but a complete response gives a genuine, specific one-sentence answer to each of the four questions (what a row represents, the time/geography covered, the suspected direction of bias, and who would be harmed by a wrong conclusion) rather than a generic or evasive answer to any of them.

***Q17.*** *Explain why answering these four questions in writing, rather than just 'thinking about it,' makes a meaningful difference in practice.*

**A:** Writing forces genuine specificity and surfaces gaps in the analyst's own understanding that remain comfortably vague when only 'thought about' — a written answer can also be checked by a supervisor or colleague, shared as documentation, and revisited later, none of which an unrecorded mental note allows.

***Q18.*** *Which of the four checklist questions in Section 1.9 do you think is most often skipped in real practice, and why?*

**A:** A reasonable answer: the 'suspected direction of bias' question is most often skipped, since it requires actively arguing against the reliability of one's own data before any analysis has even begun, which can feel counterproductive or uncomfortable compared to the more neutral, factual first two questions.

***Q19.*** *Propose one way a team lead could build this habit into a team's standard workflow, rather than relying on each individual analyst to remember it.*

**A:** A team lead could build a short provenance-check field into the team's standard project template or first-commit checklist (Chapter 6), making it a visible, required step in the project's documentation rather than an informal individual habit that varies across team members.

***Q20.*** *Explain briefly why this habit is described as something that becomes 'automatic' over time rather than remaining an effortful extra step indefinitely.*

**A:** The habit becomes automatic because, after repeated practice, the analyst begins to notice relevant provenance concerns as a natural part of first encountering any new dataset, in the same way an experienced driver checks mirrors without consciously deciding to do so each time — the underlying

vigilance becomes internalised rather than requiring deliberate, effortful recall of a checklist.

***Q21.*** *Using Section 1.10, describe a situation where waiting for a perfectly validated analysis would itself cause harm, and explain what a reasonable compromise would look like.*

**A:** An active cholera outbreak, where every day of delay could allow the outbreak to spread further, is a situation where waiting for a fully validated, triangulated analysis could itself cost lives; a reasonable compromise is acting on the best currently available evidence (e.g., a preliminary epicurve and attack rate) while explicitly flagging it as provisional and committing to update the response as more complete data becomes available.

***Q22.*** *Explain the difference between a limitation severe enough to invalidate a conclusion outright and one that can be managed with an explicit caveat, using an example.*

**A:** A severe limitation might be a dataset with no denominator information at all, making any rate calculation impossible to defend at any confidence level; a manageable limitation might be a denominator based on a slightly outdated population projection, which can be reported with an explicit caveat about its likely direction and rough size of bias rather than invalidating the entire analysis.

***Q23.*** *Why does this book argue that calibrated judgment, not a fixed rule, is the right way to balance rigour and timeliness?*

**A:** Calibrated judgment is necessary because a fixed rule (e.g., 'never report a finding unless every possible confound has been ruled out') would make timely, useful analysis impossible in practice, while a fixed rule requiring no validation at all would abandon the core discipline this book has argued for — the right balance depends on the specific stakes, time pressure, and severity of each limitation in a given situation, which cannot be reduced to a single universal rule.

***Q24.*** *Propose one way an analyst could flag 'best current evidence, still being validated' clearly to a decision-maker without either overstating or hiding the analysis's limitations.*

**A:** An analyst could use a clearly labelled 'confidence level' or 'preliminary/provisional' tag directly on a chart or report, accompanied by a one-sentence statement of the specific limitation and what would change the assessment (e.g., 'preliminary estimate; will be updated once the full register audit is complete').

***Q25.*** *How might this balance differ between a routine annual report and an active outbreak response? Explain your reasoning.*

**A:** In a routine annual report, there is generally time to pursue full triangulation and validation (as in Section 2.7's worked example) before publication; in an active outbreak response, the balance shifts substantially toward using the best available evidence quickly, with a clear commitment to revise as better data arrives, since the cost of delay itself becomes a major consideration that a routine report does not face.

***Q26.*** *Summarise, in a single paragraph, the single most important idea from Chapter 1 that you would want a brand-new research assistant to understand on their first day.*

**A:** A complete answer names a single core idea (most likely: that trusting a number requires understanding how it was generated, not just what technique was applied to it) and explains it in language a new research assistant could immediately apply to their first task.

***Q27.*** *Explain how the malaria TPR example in Section 1.5 and the IRS campaign case study in Section 1.8 both illustrate the same underlying lesson, despite being different scenarios.*

**A:** Both examples show that two people (or two data streams) looking at what appears to be 'the same' underlying phenomenon

can reach different, sometimes contradictory conclusions purely because of differences in how the data was generated or handled, not because either party made an arithmetic error — reinforcing that provenance-awareness, not just technical correctness, determines whether a conclusion is trustworthy.

***Q28.*** *Why does this chapter argue that programming skill alone, without epidemiological or provenance awareness, is not sufficient for responsible health data science?*

**A:** Programming skill alone can execute a technically correct calculation on data that is subtly wrong, biased, or inappropriate for the question being asked; without epidemiological and provenance awareness, a technically fluent analyst can produce a confidently wrong, and potentially harmful, answer just as easily as an unskilled one.

***Q29.*** *Propose one question you would ask a data provider before accepting a new dataset for analysis, based on this chapter's arguments.*

**A:** Answers will vary, but a strong example is asking directly: 'has this indicator been validated against an independent source, and if so, what was found?' — directly operationalising this chapter's core lesson.

***Q30.*** *If you had to remove one section from this chapter to make room for something else, which would you keep no matter what, and why?*

**A:** Answers will vary by individual judgement, but a well-reasoned response identifies one section (commonly Section 1.2's pipeline framework, or 1.5's worked example) as most essential to retain, with a clear justification tied to the chapter's central argument.

***Q31.*** *In the disagreement between Analyst A and Analyst B (Section 1.12), propose a third possible approach neither analyst considered, and explain its trade-offs.*

**A:** A third approach: rather than a hard exclusion threshold, weight each facility-month's contribution by its number of tests performed (giving small facilities less influence rather than excluding them entirely) — this avoids both the instability Analyst A worried about and the selection bias Analyst B worried about, at the cost of a slightly more complex analysis.

***Q32.*** *Explain why 'running the analysis both ways and checking if the conclusion changes' is a more useful response to analyst disagreement than trying to determine a single objectively correct choice.*

**A:** Running the analysis both ways directly tests whether the disagreement actually matters for the conclusion, which is the practically relevant question, whereas debating which single choice is 'more correct' in the abstract can consume time without ever resolving whether the underlying policy conclusion is robust to the choice.

***Q33.*** *Describe a disagreement you have witnessed or could imagine among colleagues analysing the same dataset, and how Section 1.12's adjudication approach would apply.*

**A:** Answers will vary by individual experience, but should describe a specific analytic choice (an exclusion rule, a variable definition, a modelling assumption) about which two reasonable people disagreed, and explain how running both versions and comparing conclusions would apply.

***Q34.*** *Why might silently picking whichever approach gives the more convenient result be especially tempting, and especially dangerous, under time pressure?*

**A:** Silently picking the more convenient result is especially tempting under time pressure because there is rarely time to run and fully investigate both versions, and it is especially dangerous because it turns an ordinary methodological judgement call into an undisclosed source of bias that could

change the substantive conclusion without anyone (including the analyst) fully recognising it.

***Q35.*** *If your reported finding depended entirely on one debatable analytic choice (such as an exclusion threshold), what should you disclose in your written report?*

**A:** A responsible report should explicitly disclose the debatable choice, briefly justify why it was made, and ideally report whether and how the conclusion would change under the alternative reasonable choice, so a reader can judge the finding's robustness for themselves.

## Chapter 2

## The Nigerian Health Data Ecosystem

***Q1.*** *For an indicator you work with regularly, identify which of the six data sources in Section 2.1 it comes from, and name one limitation specific to that source.*

**A:** Answers will vary by indicator; the key requirement is correctly matching the indicator to one of the six sources (DHIS2/NHMIS, NDHS, MICS, NCDC/SORMAS-IDSR, Global Fund malaria M&E, institutional cohort, or vital registration) and naming a limitation specific to that source (e.g., NDHS: infrequent, ~5-year gap; DHIS2: denominators depend on population projections that can drift from true catchment population).

***Q2.*** *Explain, in your own words, why an HMIS-derived test positivity rate might be biased upward relative to a register-verified value.*

**A:** HMIS-derived TPR can be biased upward because monthly summary forms are sometimes completed before all test results (particularly negative ones from later in the reporting period) are fully transcribed, and because reporting emphasis tends to favour confirmed-positive cases — both reduce the

denominator (or numerator composition) in a way that inflates the calculated positivity rate relative to a full register audit.

***Q3.*** *You are asked to estimate malaria prevalence in a state. List the sources you would triangulate and what each would contribute.*

**A:** A complete answer triangulates at least: HMIS/DHIS2 trend data (clinic-based positivity among tested febrile patients), Global Fund diagnostic register audits (a corrected version of the same clinic-based quantity), and a community prevalence survey (parasitaemia in the general population, a conceptually distinct quantity from clinic-based positivity) — explicitly noting what each source contributes and its limitation.

***Q4.*** *In the worked example of Section 2.7, explain why averaging the HMIS, register-audit, and community-survey figures would produce a meaningless number.*

**A:** Averaging would be meaningless because the three figures answer three different questions: clinic-based positivity as reported (HMIS, likely inflated), clinic-based positivity as verified (register audit, a better estimate of the same underlying quantity), and general-population prevalence regardless of care-seeking (community survey, a lower and conceptually distinct figure). Averaging a rate and a prevalence produces a number with no clear real-world referent.

***Q5.*** *A DHIS2 export arrives with one column per month (Jan_2024, Feb_2024, ...). Describe, in words, the reshaping operation needed before this data can be plotted as a time series (see Chapter 10).*

**A:** The wide format must be reshaped ('melted' or pivoted) from one row per facility with 12 month-columns into a long format with one row per facility-month and a single 'month' column and a single 'value' column, so that the data can be grouped, sorted chronologically, and plotted as a continuous time series (as required for the methods in Chapter 10).

***Q6.*** *Using the R code in Section 2.8 as a guide, write the equivalent Python (pandas) code to reshape a wide DHIS2 export into long format using pandas.melt().*

**A:** "'python import pandas as pd dhis2_wide = pd.read_csv('dhis2_export_wide.csv') dhis2_long = dhis2_wide.melt(id_vars=['facility_id'], value_vars=[c for c in dhis2_wide.columns if c.startswith('indicator_')], var_name='indicator', value_name='value') "'

***PYTHON — supporting code***

```
# Python equivalent of the Section 2.8 R reshape, using pandas.melt
import pandas as pd

dhis2_wide = pd.read_csv("dhis2_export_wide.csv")
indicator_cols = [c for c in dhis2_wide.columns if c.startswith("indicator_")]
dhis2_long = dhis2_wide.melt(id_vars=["facility_id"], value_vars=indicator_cols,
                             var_name="indicator", value_name="value")
dhis2_long["indicator"] = dhis2_long["indicator"].str.replace("indicator_", "",
regex=False)
```

***Q7.*** *Explain why NDHS individual and household recode files must be merged on cluster and household identifiers rather than simply concatenated (stacked) into one file.*

**A:** They must be merged (joined) on shared cluster and household identifiers because each file contains different variables collected at a different level (individual-level responses vs. household-level characteristics) for the same underlying households; simply stacking (concatenating) the files would incorrectly combine unrelated rows and variables rather than attaching each household's characteristics to the correct individual respondents within it.

***Q8.*** *You are told a dataset request will take '6-8 weeks for approval.' Using Section 2.9, explain how this should change your project timeline and what you could do productively while waiting.*

**A:** A 6-8 week approval timeline means data collection/access should be initiated as one of the very first project steps, well before other preparatory work is scheduled to finish, and the waiting period can be used productively to draft the data cleaning and analysis plan, prepare a data dictionary template, and pilot-test analysis code on a small simulated dataset with the same expected structure.

***Q9.*** *Identify one Nigerian health data source not explicitly listed in Section 2.1's table, and describe its unit of observation and update frequency.*

**A:** Answers will vary; acceptable examples include facility-level birth registration data (National Population Commission, individual event, continuous but under-reported), routine immunization tally sheets (facility level, monthly), or Global Fund/PEPFAR HIV programme M&E data (facility/patient level, monthly/quarterly).

***Q10.*** *Explain why an NCDC/SORMAS dataset would typically require a more restrictive access process than an NDHS dataset.*

**A:** NCDC/SORMAS data concerns active, potentially sensitive outbreak and case-level information where premature or unauthorised disclosure could affect public health response, individual privacy, or even national security/diplomatic considerations during a serious outbreak, whereas NDHS is a de-identified, aggregate-oriented survey dataset specifically designed and pre-approved for broad public health research use.

***Q11.*** *In the Section 2.10 case study, calculate the percentage difference between the highest (2.00 per 1,000) and lowest (1.43 per 1,000) rate estimates relative to the lowest estimate.*

**A:** Percentage difference = (2.00 " 1.43) / 1.43 × 100 H" 39.9%, meaning the highest denominator-based rate estimate is nearly 40% higher than the lowest — a substantial range

driven entirely by denominator choice with no change in the actual case count.

***PYTHON — supporting code***

```
# Percentage difference between the highest and lowest denominator-based rate estimates
highest, lowest = 2.00, 1.43
pct_diff = (highest - lowest) / lowest * 100
print(f"Percentage difference: {pct_diff:.1f}%")
# Percentage difference: 39.9%
```

***Q12.*** *Propose one practical step a State Ministry of Health could take to reduce the kind of denominator uncertainty illustrated in Section 2.10 for a rapidly changing LGA.*

**A:** A practical step is commissioning or supporting a rapid population verification exercise (such as a brief household listing or satellite-imagery-assisted estimate) specifically for LGAs known to have experienced significant displacement, rather than relying on extrapolated pre-displacement figures for these particular areas.

***Q13.*** *Explain why simply always using the most recent NPC projection, without further investigation, would not fully solve the denominator problem illustrated in this case study.*

**A:** Always using the most recent NPC projection would not fully solve the problem because that projection is itself based on an assumed uniform growth rate extrapolated from the last census, which does not account for the specific, unusual displacement-driven population change in this LGA — the projection method's underlying assumption is violated precisely in the cases where accurate denominators matter most.

***Q14.*** *If you had to choose just one of the three denominators in Section 2.10 to report as a headline figure, which would you choose and why, given the LGA's displacement context?*

**A:** Given the LGA's documented recent displacement, the rapid assessment figure (175,000) is likely the most currently accurate, since it is the only one of the three specifically designed to capture the post-displacement population, though this should be reported with an explicit caveat about the assessment's own margin of error.

***Q15.*** *Describe how you would flag, in a technical report, that a reported rate is sensitive to denominator choice, without simply omitting the rate altogether.*

**A:** You would state explicitly in the report's methods or footnotes which denominator source was used for each figure, note that alternative denominators exist and could shift the reported rate within a stated plausible range, and, where the choice is especially consequential (as in a displacement-affected LGA), flag this explicitly rather than presenting a single unqualified number as if it were precise.

***Q16.*** *Using the summary table in Section 2.11, identify which source you would recommend for a question about population-representative HIV prevalence, and justify your choice.*

**A:** NDHS is the recommended source for population-representative HIV prevalence because it uses a probability sample designed specifically to represent the general population, independent of who chooses to seek care or get tested, unlike facility-based sources such as DHIS2 which only capture individuals who present at a health facility.

***Q17.*** *Which source in Section 2.11's table would you recommend for near-real-time monitoring of a suspected cholera outbreak, and why?*

**A:** NCDC/SORMAS-IDSR is the recommended source for near-real-time cholera outbreak monitoring, since it is specifically designed for case-based, near-real-time reporting of priority diseases, unlike NDHS/MICS (too infrequent) or DHIS2 (aggregate, monthly, and not case-based).

***Q18.*** *Explain why 'primary strength' and 'primary caution' are both necessary parts of a useful quick-reference table, rather than strength alone.*

**A:** Strength alone is insufficient because every source's greatest strength is generally paired with a specific limitation that must be actively managed or corrected for; presenting only the strength risks an analyst using the source uncritically for a purpose it is poorly suited to, precisely the kind of naive use this book has repeatedly cautioned against.

***Q19.*** *Propose one additional column you would add to the table in Section 2.11 to make it more useful for project planning.*

**A:** A useful additional column: 'typical access process/lead time,' summarising the Section 2.9 discussion so a project planner can immediately see not just what a source offers, but how long it will realistically take to obtain access to it.

***Q20.*** *If you had to choose only one source from Section 2.11's table to fund and strengthen with limited resources, which would you choose for a state focused on malaria elimination, and why?*

**A:** For a state focused on malaria elimination, prioritising strengthening of the Global Fund malaria M&E system would likely be most valuable, since it is specifically cross-validated against source registers and directly relevant to the disease of focus, though its coverage limitation (Global Fund-supported facilities only) would need to be addressed as part of that investment.

***Q21.*** *Using Section 2.12, explain why two systems using different facility identifiers can make even a simple data merge difficult, beyond any denominator issue.*

**A:** Even with correct denominators, if System A refers to 'Eleyele PHC' by an internal numeric code and System B refers to the same facility by its full name with slightly different spelling, a simple merge on facility identifier will fail to match

the two systems' records for that facility at all, regardless of how accurate each system's own denominator or case count is.

***Q22.*** *What is HL7 FHIR, according to Section 2.12, and what problem is it intended to solve?*

**A:** HL7 FHIR (Fast Healthcare Interoperability Resources) is an international standard for exchanging clinical health data electronically in a structured, consistent format; it is intended to allow different health information systems to share and interpret clinical data about the same patient or facility consistently, rather than each system using its own incompatible internal format.

***Q23.*** *Why does the chapter describe interoperability gaps as a 'structural,' rather than a one-off data entry, problem?*

**A:** It is described as structural because the underlying cause is that different systems were designed and built independently, often by different organisations at different times, without a shared, agreed data format or identifier system from the outset — this is a design and governance issue affecting entire systems, not a one-off mistake made by an individual data entry clerk.

***Q24.*** *Propose one practical workaround a research team could use today, while waiting for full interoperability standards adoption, to merge two systems with mismatched facility names.*

**A:** A practical workaround is building and maintaining a manually curated facility-name crosswalk table (mapping each system's facility identifiers/names to a single common reference identifier), combined with fuzzy matching (Section 4.7) to catch minor spelling or formatting variations, refined and expanded over time as new mismatches are discovered.

***Q25.*** *Explain the connection between Section 2.12's interoperability problem and Section 4.7's fuzzy-matching technique.*

**A:** Section 2.12's interoperability problem is the broader, structural cause; Section 4.7's fuzzy matching is one specific, practical technique for detecting and resolving the resulting mismatches (like inconsistently typed facility names) once they occur — the fuzzy matching technique is a tactical response to the structural interoperability gap described in Section 2.12.

***Q26.*** *Explain why this chapter treats 'which data source to use' as a methodological decision on par with choosing a statistical model.*

**A:** Choosing a data source is a methodological decision because different sources answer subtly different underlying questions (as shown in Section 2.7), use different denominators, and carry different, source-specific biases — meaning the choice of source shapes the analysis's validity just as fundamentally as the choice of statistical model does.

***Q27.*** *Summarise, in your own words, the single biggest practical challenge a Nigerian health data scientist faces when combining two or more of this chapter's data sources.*

**A:** Answers will vary, but a common and defensible answer centres on reconciling incompatible denominators, time periods, or definitions across sources, which requires domain judgement rather than a purely mechanical merge.

***Q28.*** *Propose one improvement to Nigeria's health data ecosystem (realistic or aspirational) that would most reduce the kinds of problems illustrated in this chapter's case studies.*

**A:** Answers will vary; a strong response proposes something concrete and specific (e.g., a nationally standardised, regularly updated facility catchment population registry) and explains how it would directly reduce one of this chapter's documented discrepancies.

***Q29.*** *Why does this chapter emphasise triangulation over simply trusting whichever single source is easiest to access?*

**A:** Triangulation is emphasised because no single source is uniformly reliable across all contexts; relying on 'whichever is easiest to access' risks silently inheriting that source's specific, uncorrected biases without the cross-check that comparing multiple sources provides.

***Q30.*** *If you could add one new nationally-representative data source to Nigeria's health data ecosystem, what would it measure, and why?*

**A:** Answers will vary, but a strong response proposes a specific new data source, states what it would measure, and explains concretely how it would fill a gap left by the six sources already catalogued in this chapter.

***Q31.*** *Explain, using Section 2.14, how an indicator's metadata definition changing partway through a multi-year period could create an artefactual trend jump.*

**A:** If an indicator's formula changes partway through a multi-year period (for example, a new data element being added to or removed from its calculation), the resulting trend line can show an abrupt jump at exactly the point of the definition change, which a reader unaware of the change could easily mistake for a genuine shift in the underlying health outcome.

***Q32.*** *What specific check would you perform before trusting a five-year DHIS2 trend line for a single indicator?*

**A:** Before trusting a multi-year DHIS2 trend, an analyst should check the indicator's metadata version history (via the DHIS2 API or administrative interface) to confirm its defining formula remained materially unchanged throughout the period being analysed.

***Q33.*** *Why might a system administrator have a legitimate reason to change an indicator's formula, even though doing so complicates trend analysis?*

**A:** A system administrator might legitimately change an indicator's formula to correct a previously identified calculation error, to align with an updated national reporting guideline, or to incorporate a newly available data element that improves the indicator's accuracy — all legitimate reasons that nonetheless complicate historical trend comparison.

***PYTHON — supporting code***

```
# A simple facility-ID existence validation rule, as proposed in Section 2.14 / 4.13
def validate_facility_id(df, master_facility_ids):
    unknown = df.loc[~df["facility_id"].isin(master_facility_ids),
"facility_id"].unique()
    return list(unknown)

unknown_ids = validate_facility_id(sim, master_facility_ids={"F001", "F002",
"F003"})
print(f"{len(unknown_ids)} unrecognised facility ID(s) found")
```

***Q34.*** *Propose a way to flag, in a published trend chart, that an indicator's definition changed partway through the period shown.*

**A:** A reasonable approach is a footnote or annotation directly on the trend chart marking the date of any known definition change, alongside a brief note on what changed, so a reader can correctly interpret any discontinuity around that point.

***Q35.*** *How does the metadata-versioning problem in Section 2.14 relate to the data dictionary concept introduced in Section 4.8?*

**A:** Both concepts share the underlying principle that a variable's meaning must be explicitly and stably documented over time for it to be validly compared or trended — a data dictionary documents a variable's definition at a point in time, while metadata versioning tracks whether and how that definition has changed across time.

## Chapter 3

## Foundations of Probability and Statistical Thinking for Health Data

***Q1.*** *A malaria RDT has sensitivity 0.95 and specificity 0.90. If prevalence among tested patients is 40% (peak transmission season), compute the positive predictive value and compare it to the 5%-prevalence example in Section 3.4.*

**A:** Using the Bayes' theorem method: true positives = 0.95 × 400 = 380; false positives = 0.10 × 600 = 60; PPV = 380/(380+60) = 380/440 H" 86.4%. This is dramatically higher than the 5%-prevalence example's PPV of about 33.6%, illustrating that the same test is far more reliable in a high-prevalence (rainy season) setting than a low-prevalence (dry season) setting.

***PYTHON — supporting code***

```
def ppv(sensitivity, specificity, prevalence):
    tp = sensitivity * prevalence
    fp = (1 - specificity) * (1 - prevalence)
    return tp / (tp + fp)

print(f"PPV at 40% prevalence: {ppv(0.95, 0.90, 0.40):.1%}")
# PPV at 40% prevalence: 86.4%
```

***Q2.*** *Explain why monthly case counts from a single small health facility are better modelled as Poisson than as Normal.*

**A:** Small facility case counts are better modelled as Poisson because case counts are non-negative integers arising from a large number of independent low-probability opportunities for infection (a defining condition for the Poisson approximation), and because Poisson-distributed counts are inherently right-skewed with variance equal to the mean — properties a continuous, symmetric Normal distribution does not naturally capture, especially near zero.

***Q3.*** *A facility's TPR rises from 28% to 41% between two consecutive months with only 20 tests performed each month. Discuss, without doing a formal test, why this alone should not be read as evidence of a transmission increase.*

**A:** With only 20 tests per month, small-number sampling variability alone can easily produce a swing from 28% to 41% by chance (a difference of roughly 2-3 additional positive tests out of 20), so without a formal test accounting for this sample size, the swing should not be read as evidence of a real transmission increase; a much larger and more stable sample, or several consecutive months of similar magnitude, would be needed for stronger evidence.

***Q4.*** *Using the step-by-step method in Section 3.6, calculate the false negative rate (number of infected people who test negative) when prevalence is 40%, sensitivity is 0.95, and 1,000 patients are tested.*

**A:** False negatives = (1 " sensitivity) × number infected = 0.05 × 400 = 20 infected people who test negative (out of the 400 truly infected at 40% prevalence in a sample of 1,000).

***PYTHON — supporting code***

```
sensitivity, prevalence, n_tested = 0.95, 0.40, 1000
n_infected = prevalence * n_tested
false_negatives = (1 - sensitivity) * n_infected
print(f"False negatives: {false_negatives:.0f} of {n_infected:.0f} truly infected")
# False negatives: 20 of 400 truly infected
```

***Q5.*** *A new, cheaper RDT has sensitivity 0.85 and specificity 0.98. Using the Python function in Section 3.6, compute its PPV at 5% prevalence and compare it to the original test's PPV. Which property (sensitivity or specificity) matters more at low prevalence, and why?*

**A:** Using the function: PPV = (0.85 × 0.05) / [(0.85 × 0.05) + (0.02 × 0.95)] = 0.0425 / (0.0425 + 0.019) = 0.0425/0.0615

H" 69.1%, substantially higher than the original test's 33.6% PPV at the same 5% prevalence. This shows that at low prevalence, specificity matters more than sensitivity for PPV, because most tested people are truly negative, so even a small false-positive rate (1 " specificity) applied to the large negative group swamps the smaller number of true positives.

***PYTHON — supporting code***

```
def ppv(sensitivity, specificity, prevalence):
    tp = sensitivity * prevalence
    fp = (1 - specificity) * (1 - prevalence)
    return tp / (tp + fp)

cheap_test_ppv = ppv(0.85, 0.98, 0.05)
original_test_ppv = ppv(0.95, 0.90, 0.05)
print(f"Cheaper test PPV: {cheap_test_ppv:.1%}, Original test PPV:
{original_test_ppv:.1%}")
# Cheaper test PPV: 69.1%, Original test PPV: 33.6%
```

***Q6.*** *Modify the R simulation in Section 3.7 to use a Poisson rate of lambda = 3 (a low-transmission facility). Run it and comment on whether the mean and variance are still approximately equal.*

**A:** With lambda = 3, the simulated mean and variance should still both be approximately 3, since the Poisson distribution's defining property (mean = variance) holds regardless of the rate parameter's value; only the shape becomes more strongly right-skewed and concentrated near zero at this lower rate compared to lambda = 18.

***R — supporting code***

```
# R: simulating a low-transmission facility (lambda = 3) and checking mean vs
variance
set.seed(123)
low_rate <- rpois(n = 10000, lambda = 3)
cat("Mean:", mean(low_rate), " Variance:", var(low_rate), "\n")
# Mean and variance should both be close to 3
```

***Q7.*** *In your own words, explain the difference between P(Disease | Positive test) and P(Positive test | Disease), and why confusing the two is a common and consequential error in interpreting diagnostic tests.*

**A:** P(Disease | Positive test) is the probability a patient actually has the disease given they tested positive (what a clinician wants to know — the positive predictive value); P(Positive test | Disease) is the probability a diseased patient tests positive (the test's sensitivity, a fixed property of the test). Confusing the two ('the test is 95% sensitive, so a positive result means 95% chance of disease') ignores the base rate (prevalence) entirely and can badly overstate the reliability of a positive result in a low-prevalence population, as shown throughout this chapter.

***Q8.*** *Give one example, other than those in Section 3.8, of health data where treating each row as independent would be incorrect, and explain the true unit of independence.*

**A:** Example: repeated blood pressure measurements taken from the same patient at multiple clinic visits are not independent of each other, since they share the patient's underlying, relatively stable physiology; the true unit of independence here is the patient, not the individual visit-level measurement.

***Q9.*** *Explain why the variance of a sum of two correlated random variables is not simply the sum of their individual variances.*

**A:** The variance of a sum of correlated variables includes an additional covariance term (positive if the variables are positively correlated), so simply summing the individual variances (as would be correct only under independence) omits this term and gives an incorrect (typically too small, for positively correlated variables) total variance.

***Q10.*** *A colleague computes a standard error assuming independence for data that is actually clustered by facility. Will this standard error typically be too large or too small? Explain why.*

**A:** The naive standard error will typically be too small, because it fails to account for the fact that observations within the same facility are more similar to each other than independent observations would be, meaning there is effectively less independent information in the dataset than the raw sample size suggests — this is directly analogous to the survey design effect discussed in Chapter 7.

***Q11.*** *Using the pharmacovigilance example in Section 3.9, recompute the positive predictive value if the true prevalence of the adverse reaction were 2% rather than 0.5%, keeping sensitivity and specificity unchanged.*

**A:** Using the same PPV formula with prevalence = 0.02: true positives = 0.90 × 0.02 = 0.018; false positives = 0.05 × 0.98 = 0.049; PPV = 0.018/(0.018+0.049) H" 26.9% — considerably higher than the 8.3% PPV at 0.5% prevalence, illustrating how much PPV improves even at a still-fairly-low absolute prevalence increase.

***PYTHON — supporting code***

```
def ppv(sensitivity, specificity, prevalence):
    tp = sensitivity * prevalence
    fp = (1 - specificity) * (1 - prevalence)
    return tp / (tp + fp)

print(f"PPV at 2% prevalence: {ppv(0.90, 0.95, 0.02):.1%}")
# PPV at 2% prevalence: 26.9%
```

***Q12.*** *Explain why a 90% sensitivity, 95% specificity screening tool can still have a very low positive predictive value, without this being a sign the tool is poorly built.*

**A:** A low PPV despite strong sensitivity and specificity is a mathematical consequence of screening for a rare event in a large population — most of the population is truly negative, so even a small false-positive rate applied to that large negative

group produces more false positives than true positives, a base-rate effect entirely independent of how well-built the tool is.

***Q13.*** *Propose one way a pharmacovigilance team could communicate the tool's output to clinical staff so that a 'flagged' record is not mistaken for a confirmed adverse reaction.*

**A:** The team could explicitly label flagged records as 'recommended for manual chart review' rather than 'confirmed adverse reaction,' and could accompany any performance reporting with the PPV at the tool's actual deployment prevalence, not just its sensitivity and specificity, so clinical staff calibrate their expectations correctly.

***Q14.*** *In Section 3.9, explain why increasing specificity would generally help the tool's positive predictive value more than a similar increase in sensitivity, given the very low base rate.*

**A:** Because the false-positive group (calculated from the very large truly-negative population) dominates the denominator of the PPV calculation at low prevalence, reducing the false-positive rate (increasing specificity) has a proportionally larger effect on PPV than increasing sensitivity, which only affects the smaller true-positive numerator.

***Q15.*** *Give one other Nigerian health screening context (besides malaria RDTs and pharmacovigilance) where the base rate fallacy could mislead a non-technical audience, and briefly explain why.*

**A:** One example: a national screening algorithm flagging patients as at high risk of a rare cancer from a large administrative claims dataset — even with strong sensitivity and specificity, the very low disease prevalence in the general screened population would produce a low PPV, and reporting only the sensitivity/specificity figures to the public could create unrealistic expectations that a flagged result is a near-certain diagnosis.

***Q16.*** *Explain, using Section 3.10, why a very large sample drawn only from urban, private-hospital patients would not necessarily give a trustworthy estimate of national child mortality.*

**A:** A very large sample drawn only from urban, private-hospital patients would converge, with great statistical precision, toward the mortality experience of that specific, systematically wealthier and better-resourced subpopulation — not toward the true national child mortality rate, which includes many rural and lower-income households the sample never touches.

***Q17.*** *What does the law of large numbers guarantee, precisely, and what does it NOT guarantee, according to Section 3.10?*

**A:** The law of large numbers guarantees that a sample average converges toward whatever true value the SAMPLING PROCESS ITSELF is actually drawing from as sample size grows; it does not guarantee that this value equals the broader population parameter of genuine interest, unless the sampling process is representative of that broader population.

***Q18.*** *Give an example of a Nigerian health dataset that could be very large in sample size but still systematically biased in the way Section 3.10 describes.*

**A:** Example: a very large dataset of mobile phone-based health app users, which could easily include hundreds of thousands of records, would still systematically exclude anyone without a smartphone or reliable internet access — likely a substantial, poorer, and more rural portion of the population — regardless of how large the raw sample size grows.

***Q19.*** *Why is 'we have a huge sample size' sometimes used inappropriately to defend a conclusion, in light of Section 3.10?*

**A:** This claim is sometimes used inappropriately to imply the estimate must be reliable, when in fact a large sample size only protects against random sampling error, not against a systematic

bias baked into how the sample was selected in the first place — the two protections are entirely separate.

***Q20.*** *Explain briefly how you would investigate whether a large dataset's size is actually protecting it from bias, rather than simply assuming it does.*

**A:** You would examine how the dataset's participants/records were actually recruited or selected, compare basic demographic characteristics against known, independently-sourced population benchmarks (e.g., from a census or a representative survey), and look for evidence that the sampling process systematically favours or excludes any particular subgroup relevant to the outcome being studied.

***Q21.*** *Using Section 3.11, explain why some improvement would be expected at follow-up even if an intervention targeting the worst-performing facilities had no true effect at all.*

**A:** Even with a genuinely ineffective intervention, facilities selected specifically because they had an unusually high baseline TPR (partly due to chance factors in that particular baseline period) would, on average, show a somewhat lower TPR at the next measurement simply because extreme values tend to be followed by less extreme values — this improvement would occur even with no true underlying change in transmission or care quality.

***Q22.*** *How is regression to the mean different from the seasonal and reporting-completeness confounds discussed in Section 1.8?*

**A:** Regression to the mean arises purely from the statistical property of selecting on an extreme value and does not require any change in reporting completeness or season at all; the confounds in Section 1.8 instead involve genuine changes in the underlying data-generating process (who is reporting, and what time of year it is) rather than a purely statistical artefact of selecting extreme baseline observations.

***Q23.*** *Propose a study design feature that could help distinguish a genuine intervention effect from regression to the mean.*

**A:** Including a comparison group of facilities NOT selected for the intervention (but with similarly extreme baseline values) allows an evaluator to observe how much 'natural' improvement occurs in similarly extreme facilities without any intervention at all, isolating the genuine intervention effect as the DIFFERENCE between the treated and comparison groups' improvement, rather than attributing all improvement in the treated group to the intervention.

***Q24.*** *Why does selecting facilities specifically BECAUSE of an extreme baseline value make regression to the mean especially likely to occur?*

**A:** Selecting on an extreme baseline value specifically targets facilities or patients whose extreme reading was, at least in part, a product of chance variation around their own true underlying average; because chance is unlikely to repeat in the same extreme direction at the next measurement, some reversion toward a more typical value is a near-mathematical certainty for this specifically-selected group, regardless of any true underlying trend.

***Q25.*** *Explain, in your own words, why regression to the mean is a genuine statistical phenomenon and not evidence of dishonest reporting.*

**A:** Regression to the mean is a genuine statistical phenomenon following directly from the role of random variation in producing any single extreme observation — it is not evidence of dishonest reporting or manipulation, and understanding it is meant to help an evaluator correctly attribute observed improvement between a genuine intervention effect and this expected, purely statistical reversion, not to cast doubt on the honesty of the underlying data.

***Q26.*** *Explain, in a few sentences, why this chapter argues probability must be understood before statistical inference can be interpreted correctly.*

**A:** Every statistical inference technique (confidence intervals, hypothesis tests, model-based predictions) is a probability statement about an assumed underlying process; without understanding that underlying probability model, a practitioner can compute a correct-looking number whose actual meaning they do not truly grasp.

***Q27.*** *Summarise the common thread linking the diagnostic testing example (Section 3.4), the pharmacovigilance example (Section 3.9), and the law of large numbers discussion (Section 3.10).*

**A:** The common thread is that human intuition about probability is systematically unreliable in predictable ways — whether via the base rate fallacy in diagnostic and pharmacovigilance contexts, or via a mistaken belief that small samples should closely resemble long-run averages (a naive misapplication of the law of large numbers) — and formal probability reasoning is needed to correct for this.

***Q28.*** *Why does this chapter emphasise working through calculations by hand at least once, rather than relying solely on software output?*

**A:** Working through a calculation by hand forces the practitioner to engage with each step of the underlying logic explicitly, building an intuitive grasp of why the result comes out as it does, in a way that simply reading a software-generated number does not provide.

***Q29.*** *Propose a scenario from your own field where the base rate fallacy could mislead a well-intentioned analyst.*

**A:** Answers will vary, but a strong example involves a rare outcome (e.g., a rare adverse drug reaction, or a rare disease)

where an alarming-sounding raw count or percentage could mislead a reader who does not consider the low underlying base rate.

***Q30.*** *What single probability concept from this chapter do you find most useful for your own work, and why?*

**A:** Answers will vary by individual, but a complete answer names one specific concept from the chapter (e.g., Bayes' theorem, or independence) and explains, with a concrete example from the respondent's own work, why it is the most immediately useful.

***Q31.*** *Describe, using Section 3.13, a scenario in clinical decision-making where the gambler's fallacy could subtly bias a diagnosis.*

**A:** A clinician who has just seen several consecutive negative results in a busy fever clinic might unconsciously become more inclined to diagnose the next patient as malaria-positive 'because a positive is due,' even though the true probability of that specific patient having malaria depends only on their own individual clinical presentation and the underlying prevalence, not on the sequence of previous patients' results.

***Q32.*** *Why is naming a cognitive bias explicitly (rather than leaving it as an abstract possibility) helpful for catching it in real time?*

**A:** Naming a bias explicitly, with a specific label and a memorable description, gives a practitioner a concrete mental trigger to recognise the pattern of thought as it is happening ('this is the gambler's fallacy'), which is considerably easier to catch in the moment than trying to notice an abstract, unnamed lapse in reasoning.

***Q33.*** *Explain why three consecutive negative test results do not make a positive result 'due' on the fourth test, assuming a stable underlying prevalence.*

**A:** Assuming test results are independent (a reasonable assumption when each patient's true infection status is not influenced by other patients' test outcomes), each test's result provides no information about the probability of the next test's result, so a fourth negative result remains just as likely as the third, regardless of how many negatives preceded it.

***Q34.*** *Propose one other context (outside diagnostic testing) where a health professional might unconsciously apply the gambler's fallacy.*

**A:** Answers will vary, but a plausible example is an emergency triage nurse becoming more inclined to flag the next patient as high-acuity after several consecutive low-acuity presentations, reasoning that 'a serious case is due,' despite each patient's actual acuity being independent of the sequence before them.

***Q35.*** *How does the gambler's fallacy relate to, but differ from, the base rate fallacy discussed earlier in this chapter (Section 3.9)?*

**A:** The base rate fallacy involves ignoring the underlying prevalence (base rate) of a condition when interpreting a single test result, while the gambler's fallacy involves incorrectly believing that a sequence of independent past outcomes changes the probability of a future outcome — both are errors of probabilistic reasoning, but they operate on different aspects of the underlying probability model.

## Chapter 4

## Data Management

### Types, Sources, Collection, and Cleaning

***Q1.*** *Classify each of the following as MCAR, MAR, or MNAR, with justification: (a) a laboratory machine breakdown causes random missing test results for one week; (b) sicker patients are more likely to miss a follow-up visit; (c) younger respondents*

*are more likely to skip an income question, but income itself doesn't predict skipping once age is accounted for.*

**A:** (a) MCAR — the machine breakdown is unrelated to any patient or test characteristic, a purely random equipment failure. (b) MNAR — missingness (missing the follow-up visit) is directly related to the unobserved outcome itself (how sick the patient is). (c) MAR — missingness depends on an observed variable (age) but not on the unobserved income value itself once age is accounted for.

***PYTHON — supporting code***

```
# A minimal MCAR/MAR/MNAR classification helper for exercise 1's three
scenarios
scenarios = {
   "lab_breakdown": "MCAR",       # unrelated to any variable
   "sicker_miss_visit": "MNAR",   # depends on the unobserved outcome itself
   "younger_skip_income": "MAR",  # depends on observed age, not the income
value itself
}
for scenario, mechanism in scenarios.items():
   print(f"{scenario}: {mechanism}")
```

***Q2.*** *Write out (in pseudocode or R/Python) a cleaning check that would have caught a 'pregnant male' record in a health survey dataset.*

**A:** "'r # Flag implausible sex/pregnancy combinations flag <- data$sex == 'male' & data$pregnant == 'yes' sum(flag, na.rm = TRUE) '"ThIS simple logical check cross-references two fields that should never both be true, flagging any row where they are.

***Q3.*** *Explain why LOCF (last observation carried forward) can bias a longitudinal analysis of a deteriorating clinical condition.*

**A:** LOCF assumes a patient's last observed value remains constant going forward, but in a deteriorating condition, the

true value is likely worsening after the last observation — so LOCF will systematically understate the severity of decline, biasing the analysis toward showing less deterioration (or a falsely stable/improving trend) than actually occurred.

***Q4.*** *In the worked example of Section 4.6, explain in your own words why a sensitivity analysis would be needed if the missingness mechanism were MNAR rather than MAR.*

**A:** Under MAR, multiple imputation correctly models the missing follow-up scores using observed predictors (baseline score, food security, etc.), producing unbiased estimates because all the information driving missingness is captured in the model. Under MNAR, the missingness depends on the unobserved follow-up score itself (e.g., the sickest adolescents specifically not returning) — no amount of conditioning on observed variables can fully correct this, so a sensitivity analysis testing a range of plausible assumptions about the unobserved values is needed to assess how much the conclusions might change.

***Q5.*** *Using the R code in Section 4.6 as a template, describe (in words) what would change in the 'mice()' call if the outcome variable itself (not just the predictors) had missing values in some imputed datasets but not others.*

**A:** If the outcome variable itself has missing values, 'mice()' would need to include it in its own imputation model (typically it is by default, since 'mice' imputes all variables with missingness), and the analyst must ensure the imputation model for the outcome includes all variables that will later be used as predictors of it, to avoid a subtly inconsistent (uncongenial) imputation model.

***Q6.*** *Run the Python fuzzy-matching code in Section 4.7 (or describe its logic if you cannot run it) on a list of five deliberately misspelled facility names you invent, and report which pairs are flagged.*

**A:** Answers will vary based on invented names, but should demonstrate that ‘fuzz.ratio()‘ returns a high similarity score (typically well above 90) for near-identical strings differing only by abbreviation, spacing, or minor spelling differences (e.g., ‘Eleyele PHC’ vs. ‘Eleyele Primary Health Center’), while dissimilar names score much lower.

***PYTHON — supporting code***

```
from rapidfuzz import fuzz

names = ["Eleyele PHC", "Eleyele Primary Health Cente", "Bodija Clinic",
        "Bodija Health Clinic", "Sango Health Post"]
pairs = [(a, b, fuzz.ratio(a, b)) for i, a in enumerate(names)
        for b in names[i+1:] if fuzz.ratio(a, b) > 80]
for a, b, score in pairs:
   print(f"{a!r} vs {b!r}: similarity {score:.0f}")
```

***Q7.*** *A supervisor asks why you spent two extra days building a scripted cleaning pipeline instead of ‘just fixing it in Excel.’ Write a two- to three-sentence response drawing on the reproducibility arguments of Section 4.4.*

**A:** A strong response notes that a scripted pipeline is auditable (every step is documented and can be checked), reversible (raw data is never overwritten, so mistakes can be undone), and reusable (the same script re-runs correctly next month on updated data), whereas manual Excel edits leave no record of what was changed, cannot be verified by a reviewer, and must be repeated (with the same risk of inconsistency) every reporting cycle.

***Q8.*** *Using Section 4.8, draft three rows of a data dictionary (variable, definition, type, allowed range) for a dataset you work with regularly.*

**A:** Answers will vary, but a complete row includes a variable name, a plain-language definition specific enough to remove ambiguity, a data type, and a stated plausible range or set of

allowed values (e.g., 'patient_age_years: age of patient in completed years at time of test; integer; range 0-110').

***Q9.*** *Explain why an auto-generated data dictionary skeleton (Section 4.8's R code) still requires manual completion rather than being used as-is.*

**A:** An auto-generated skeleton only captures what can be inferred mechanically from the data itself (variable names, types, an example value, missingness counts) — it cannot capture the actual meaning, units, or provenance of a variable, which requires human knowledge of how and why the data was collected, so the analyst must still manually complete the definition, units, and source fields.

***Q10.*** *A dataset arrives with no documentation at all. Rank, in order of priority, the first three things you would do before beginning any analysis.*

**A:** A reasonable priority order: (1) identify and, where possible, contact the person or team who generated the data to obtain basic documentation or clarify variable meanings; (2) perform the basic structural inspection from Chapter 1 (dimensions, types, plausibility) to form an initial impression of data quality; (3) build a data dictionary skeleton and begin filling it in based on whatever documentation or institutional knowledge can be gathered, flagging any variable whose meaning remains genuinely unclear.

***Q11.*** *In the Section 4.9 case study, explain what specific piece of information the second (reproducible) team could produce in minutes that the first team could not reconstruct at all.*

**A:** The second team could reproduce the exact excluded-record count (161, broken into four documented categories) in minutes directly from their version-controlled cleaning script, whereas the first team had no record at all of which specific records were excluded or why, since the exclusions were applied manually and never documented.

***Q12.*** *Why is 're-running the analysis under an alternative exclusion policy' a valuable capability for responding to a journal reviewer's request, beyond just defending the original figure?*

**A:** Beyond simply defending the original figure, the ability to re-run under an alternative policy lets the team directly demonstrate to a skeptical reviewer how robust (or sensitive) the paper's findings are to a specific alternative analytic choice, which is a much stronger form of evidence than merely asserting that the original choice was reasonable.

***Q13.*** *Describe, in your own words, why manual point-and-click filtering in statistical software creates the same reproducibility risk as manual spreadsheet editing, even though no spreadsheet is involved.*

**A:** Manual point-and-click filtering leaves the same invisible, unrecoverable trail as manual spreadsheet editing — even though no spreadsheet file is directly edited, the sequence of clicks and menu selections applied is not captured anywhere, so neither the original analyst nor anyone else can later reconstruct exactly what was done, for the same underlying reason: the process, not just the final data file, was never recorded.

***Q14.*** *If you inherited the first (non-reproducible) team's dataset and raw files today, what would be your first practical step toward reconstructing a defensible cleaning trail?*

**A:** The first practical step would be to gather and interview the original team members (if still available) while their memory of the process is freshest, cross-referencing their recollections against any available raw and 'final' dataset versions to reverse-engineer the most likely sequence of exclusions, before urgently establishing a version-controlled, scripted process for all future work on this dataset.

***Q15.*** *Explain why 'we are confident the original exclusions were reasonable' is not, by itself, a sufficient response to a reviewer's request to see the exact cleaning steps.*

**A:** This is insufficient because a reviewer's concern is not primarily about whether the original judgment was reasonable in principle, but about verifiability — the reviewer (or a future replication attempt) needs to be able to see and independently assess the exact steps taken, and 'we are confident it was fine' provides no way to do this.

***Q16.*** *Using Section 4.10, explain the difference between listwise deletion and winsorising, and give a scenario where each would be more appropriate.*

**A:** Listwise deletion removes an entire record if any of its variables is missing, appropriate mainly when missingness is rare and MCAR (Chapter 4); winsorising instead caps extreme but present values at a chosen percentile rather than deleting the record, appropriate when a value is present but implausibly extreme and the rest of that record's information is still valuable to retain.

***Q17.*** *Why is 'data lineage' (Section 4.10) a stronger standard than simply keeping the raw and cleaned files in the same folder?*

**A:** Data lineage documents the exact sequence of transformations applied to derive each specific value, which is a stronger and more specific standard than simply keeping raw and cleaned files in the same folder, since the latter shows the starting and ending points but not the auditable trail of exactly what happened in between, which is what a reviewer or future analyst actually needs to verify or reproduce the work.

***Q18.*** *Explain hot-deck imputation in your own words, and describe one advantage multiple imputation (Section 4.6) has over it.*

**A:** Hot-deck imputation fills a missing value by copying an observed value from a similar record elsewhere in the same dataset; multiple imputation improves on this by generating several plausible values (reflecting genuine uncertainty about the true missing value) rather than a single deterministic substitute, and by explicitly propagating that uncertainty into the final standard errors, which hot-deck imputation does not naturally do.

***Q19.*** *Propose a deduplication key (Section 4.10) appropriate for a facility-month malaria dataset, and explain why you chose those specific variables.*

**A:** A reasonable deduplication key for a facility-month dataset: facility_id combined with report_month, since together these two variables should uniquely identify a single, non-duplicated record — any pair of rows sharing both values very likely represents the same underlying facility-month observation entered more than once.

***Q20.*** *Why might winsorising be preferred over deleting outlier records entirely, in a dataset where sample size is already limited?*

**A:** In a dataset with limited sample size, deleting outlier records entirely discards potentially valuable information from the rest of that record's variables; winsorising instead retains the record (with its extreme value adjusted to a defensible bound) preserving statistical power for the record's other variables while still limiting the outlier's disproportionate influence on the specific variable in question.

***Q21.*** *Using Section 4.11, list three pieces of dataset-level metadata that a variable-level data dictionary (Section 4.8) alone would not capture.*

**A:** Three examples: (1) who collected the data and under what ethical approval; (2) the exact geographic scope and date range

covered; (3) the sampling or census methodology used to select participants or facilities — none of which is captured by a variable-level dictionary describing individual column definitions.

***Q22.*** *Why can a perfectly clean, well-structured dataset still be misapplied without proper metadata, according to Section 4.11?*

**A:** A dataset can be misapplied without proper metadata because its structure alone (variable names, types, ranges) does not reveal WHO or WHAT POPULATION it actually represents, or under what conditions it was collected — a later user could reasonably but incorrectly assume the dataset is more broadly representative, more recently collected, or more completely validated than it actually is, without dataset-level metadata to correct that assumption.

***Q23.*** *Propose a real or plausible scenario where missing dataset-level metadata led (or could lead) to a dataset being used for a population or purpose it was not designed for.*

**A:** A plausible scenario: a dataset originally collected only from patients presenting at a single tertiary referral hospital is later used, without adequate metadata clarifying this narrow scope, to estimate a disease's general population prevalence, producing a substantially biased and misleading estimate because the dataset's true (narrow, hospital-based) scope was not clearly documented and communicated.

***Q24.*** *Why is 'known, unresolved data quality concerns at the time of release' an important item to include in dataset metadata?*

**A:** This item is important because a later user needs to know not just what the data IS, but what its known limitations WERE at the time it was released, so they can make an informed judgment about whether those limitations matter for their specific intended use — omitting this information effectively

hides relevant caveats from anyone using the dataset after the original team's direct involvement has ended.

***Q25.*** *How would you store and share this dataset-level metadata alongside a dataset in your own work?*

**A:** A reasonable practice is maintaining a single 'README' or metadata file alongside the dataset itself (in the same shared repository or folder, per Chapter 6's project structure), covering all the dataset-level items listed in Section 4.11, updated whenever the dataset itself is updated or re-shared with a new collaborator.

***Q26.*** *Summarise, in your own words, why this chapter treats data cleaning as inseparable from data ethics and governance rather than a purely technical step.*

**A:** Data cleaning decisions embed ethical and governance choices (what counts as an error worth correcting, how missingness is handled, who is accountable for a cleaning decision) that cannot be separated from the technical mechanics of the cleaning process itself, which is why this chapter treats them as one integrated topic rather than two separate ones.

***Q27.*** *Explain the common thread linking missing data handling (Section 4.3), duplicate detection (Section 4.4), and metadata documentation (Section 4.11).*

**A:** The common thread is that every stage of data preparation — deciding how to treat a missing value, identifying a duplicate, or documenting what a dataset actually contains beyond its individual variables — requires the same underlying discipline of making an explicit, documented, defensible decision rather than an invisible, ad hoc one.

***Q28.*** *Why does this chapter insist on scripted rather than manual cleaning, even when manual editing might seem faster for a small dataset?*

**A:** Manual editing leaves no reproducible record of what was changed and why, making it impossible for anyone (including

the original analyst, later) to verify, audit, or correct the cleaning process, whereas a scripted approach — even if marginally slower for a very small dataset — remains fully auditable and reusable.

***Q29.*** *Propose one cleaning habit from this chapter you intend to adopt in your own work, and explain why you chose it.*

**A:** Answers will vary by individual, but should name one specific concrete habit (e.g., 'always producing a data dictionary before beginning analysis') and explain why it was chosen over the chapter's other recommended habits.

***Q30.*** *If a colleague asked you to summarise this entire chapter in one sentence, what would you say?*

**A:** Answers will vary, but a strong one-sentence summary captures the idea that data cleaning is a documented, ethically-aware, reproducible process — not a private, informal step that happens before the 'real' analysis begins.

***Q31.*** *Explain the difference between the automated validation rules in Section 4.13 and the one-off cleaning scripts introduced earlier in this chapter.*

**A:** Automated validation rules run continuously at the point of new data entry or submission, catching problems immediately and before flawed data enters the analytic dataset at all, whereas the one-off cleaning scripts described earlier are typically run after the fact, on a dataset that has already been assembled, identifying and correcting (or flagging) problems that have already occurred.

***Q32.*** *Why is preventing a data quality problem at the point of entry generally preferable to catching it downstream during cleaning?*

**A:** Preventing a problem at the point of entry is generally preferable because it stops the error from propagating into any

downstream use of the data (multiple analyses, dashboards, or reports), whereas catching it later during cleaning means the error may already have influenced decisions made before the cleaning step was performed.

***Q33.*** *Using the Python code in Section 4.13, describe what additional validation rule you would add for a dataset you work with, and why.*

**A:** Answers will vary, but a reasonable additional rule for many facility-level datasets would check that the reporting facility ID actually exists in a maintained master facility list, flagging any submission from an unrecognised or mistyped facility identifier.

***PYTHON — supporting code***

```
def validate_facility_exists(row, master_list):
    return row["facility_id"] in master_list

# Applied row-wise as an additional validation rule alongside Section 4.13's
checks
```

***Q34.*** *What is a potential downside of an overly aggressive automated validation rule (for example, one that rejects too many legitimate submissions)?*

**A:** An overly aggressive validation rule risks rejecting legitimate but unusual submissions (for example, a facility with a genuine, real outbreak-driven surge in testing volume), creating a data quality problem of its own — either lost legitimate data or frustrated, mistrustful data entry staff who begin working around the validation system.

***Q35.*** *How does DHIS2's support for configurable validation rules relate to the interoperability standards discussed elsewhere in this book (Section 2.12)?*

**A:** Both reflect the broader principle that health data systems function better when data quality and format expectations are defined and enforced systematically (via configurable rules or agreed standards) rather than being handled inconsistently and informally by whichever analyst happens to receive the data downstream.

## Chapter 5

### Exploratory Data Analysis and Visualization

***Q1.*** *Given a facility-level dataset with monthly TPR, propose one univariate, one bivariate, and one multivariate exploratory plot, and state what each is intended to reveal.*

**A:** Univariate: a histogram of monthly TPR values across all facility-months, to check its distribution and identify implausible spikes. Bivariate: a scatterplot of TPR against rainfall (or a boxplot of TPR by season), to explore the seasonal relationship. Multivariate: a scatterplot of TPR over time, faceted or coloured by state, to see whether the seasonal pattern is consistent across states or differs regionally.

***Q2.*** *A news report shows a bar chart of vaccination coverage with the y-axis starting at 70% rather than 0%. Explain how this could mislead a reader, using a concrete numeric example.*

**A:** Truncating the y-axis at 70% instead of 0% visually exaggerates the apparent size of any change: for example, a rise from 72% to 76% coverage (a 4 percentage-point, or about 5.6% relative, increase) would visually appear to nearly double the bar's height on a 70-100% axis, compared to the barely perceptible change it would show on a full 0-100% axis — misleading a reader into overestimating the magnitude of improvement.

***PYTHON — supporting code***

```
coverage_before, coverage_after = 70, 76
change_pp = coverage_after - coverage_before
relative_change = change_pp / coverage_before * 100
print(f"Absolute change: {change_pp} pp, Relative change:
{relative_change:.1f}%")
# On a 70-100 axis this modest 4pp / 5.6% relative change appears visually
exaggerated
```

***Q3.*** *Why should a relationship discovered by scanning 30 variables for correlations with an outcome be treated differently from a single pre-specified hypothesis test?*

**A:** A relationship found by scanning 30 variables faces a much higher chance of a false positive purely by chance (the multiple comparisons problem, formalised in Chapter 7) than a single, theory-driven, pre-specified hypothesis test; the scanned relationship should be treated as a candidate to be confirmed on independent data, not reported with the same confidence as a planned test.

***Q4.*** *In the worked example of Section 5.6, explain why the median and IQR were chosen over the mean and standard deviation for the depressive symptom score.*

**A:** The median and IQR were chosen because the symptom score distribution was right-skewed (most adolescents have low scores, with a long tail of severely affected individuals) — in a skewed distribution, the mean is pulled toward the tail and does not represent a 'typical' value well, whereas the median and IQR are robust to this skew and better describe the central tendency and spread actually experienced by most adolescents.

***Q5.*** *A report states 'malaria cases in LGA X rose 33% this month.' You learn the actual counts were 3 cases last month and 4 this month. Write two sentences explaining why the percentage figure alone is misleading here.*

**A:** With such small base counts, a 'rise' from 3 to 4 cases is well within the range of ordinary random month-to-month fluctuation and could easily reverse next month; presenting it as a 33% increase, without the underlying counts, makes a statistically unremarkable fluctuation sound like a substantial and concerning trend.

### *PYTHON — supporting code*

```
cases_last_month, cases_this_month = 3, 4
pct_change = (cases_this_month - cases_last_month) / cases_last_month * 100
print(f"Percentage change: {pct_change:.0f}% (raw counts: {cases_last_month} ->
{cases_this_month})")
# Percentage change: 33% -- technically correct but a trivial base-count fluctuation
```

***Q6.*** *Using the R code in Section 5.6, describe what change in the boxplot you would expect if severely food-insecure adolescents had, in reality, both a higher median symptom score and much greater variability than the other groups.*

**A:** If severely food-insecure adolescents had both a higher median AND much greater variability, the boxplot for that group would show its box (and whiskers) shifted upward (higher median line) and also stretched taller/wider (larger interquartile range and longer whiskers) compared to the other groups' more compact boxes.

### *R — supporting code*

```
# R: simulating the boxplot pattern described in exercise 6
library(ggplot2)
set.seed(1)
sim_df <- data.frame(
  food_security = rep(c("secure", "mild", "severe"), each = 100),
  score = c(rnorm(100, 5, 2), rnorm(100, 7, 2.5), rnorm(100, 11, 4))  # severe:
higher mean AND spread
)
ggplot(sim_df, aes(x = food_security, y = score, fill = food_security)) +
geom_boxplot()
```

***Q7.*** *List two of the EDA pitfalls in Section 5.7 that you have personally observed (or could imagine occurring) in a report you have read or written, and explain briefly how you would correct each.*

**A:** Answers will vary, but should identify at least two specific pitfalls (e.g., reporting a mean for a skewed variable, or treating a single scanned correlation as a confirmed finding) and describe a concrete correction (e.g., switching to median/IQR, or explicitly labelling the finding as a hypothesis requiring confirmation).

***Q8.*** *Explain why a small-multiples plot (Section 5.8) would be more useful than a single overlaid line chart for comparing TPR trends across all 36 states plus the FCT.*

**A:** A small-multiples plot would be more useful because a single chart with 37 overlapping lines becomes visually unreadable — impossible to trace any individual state's trend or spot an outlier — whereas a grid of 37 small, identically-scaled panels allows the reader to scan quickly across states for the general pattern while still being able to identify any state whose trend clearly diverges from the rest.

***Q9.*** *Propose one dataset from your own work where a small-multiples approach would be more informative than a single combined chart, and explain why.*

**A:** Answers will vary, but a complete answer names a specific dataset with many groups (e.g., facility-level trends across dozens of facilities) and explains that a combined chart would be unreadable at that scale, whereas small multiples would preserve individual-group detail.

***Q10.*** *What is one disadvantage of small multiples compared to a single combined chart, particularly when comparing exact values between two specific groups?*

**A:** A disadvantage is that comparing exact values between two specific groups is harder in small multiples than in a single overlaid chart, since the reader must visually cross-reference two separate small panels rather than directly comparing two lines occupying the same axes; a combined chart (perhaps limited to just the two groups of interest) would be better suited to that specific comparison task.

***Q11.*** *In the Section 5.9 case study, explain why looking at the distribution of individual item-level responses (not just the mean) at the low-scoring site was the key diagnostic step.*

**A:** Looking at the full item-level response distribution, rather than just the summary mean, revealed that responses were suspiciously concentrated at the lowest response option across nearly every item — a pattern much more consistent with a systematic measurement or translation artefact than with a genuine, evenly-distributed reduction in symptom severity across all items, which a lower mean alone could not distinguish from a genuine finding.

***Q12.*** *Propose one additional check, beyond the item-level distribution, that could have helped confirm or rule out the translation-quality explanation earlier.*

**A:** An additional useful check would be comparing inter-item correlations (internal consistency) at the low-scoring site against the other sites — a translation or administration problem would often also disrupt the expected pattern of correlation between related symptom items, providing further corroborating evidence beyond the marginal distribution shift alone.

***Q13.*** *Explain why jumping directly to a regression model comparing sites, without first doing the exploratory work in Section 5.9, would likely have missed this problem entirely.*

**A:** Jumping directly to a regression model comparing sites would likely have produced a statistically significant 'site'

coefficient, which could easily have been reported and interpreted as a genuine finding about geographic variation in adolescent mental health, since a regression model on its own has no mechanism for questioning whether the underlying measurement is valid at each site — that scrutiny only comes from the kind of exploratory investigation described in Section 5.9.

***Q14.*** *If the low-scoring site's item-level distribution had instead looked similar in shape to the other sites' distributions (just shifted lower), would this change your interpretation? Explain why or why not.*

**A:** If the site's distribution had the same overall shape (not concentrated at the floor) but was simply shifted uniformly lower, this would be somewhat less immediately suggestive of a measurement artefact and would warrant more (though not necessarily different) investigation — a uniform shift is at least superficially more consistent with a genuine, real difference than a distribution artificially piled up at one extreme response option.

***Q15.*** *Describe how you would report this discovered instrument-translation issue in a study's limitations section, in two to three sentences.*

**A:** A reasonable limitations statement: 'A subsequent review found that participants at one site completed a locally back-translated version of the symptom instrument that had not undergone the same formal validation as the instrument used at other sites; results for this site should therefore be interpreted with caution and are excluded from the primary cross-site comparison.'

***Q16.*** *Using the checklist in Section 5.10, review a chart you have seen recently (in a report, news article, or presentation) and identify one question from the checklist it fails.*

**A:** Answers will vary by chosen chart, but a complete response identifies a SPECIFIC checklist item the chart fails (e.g., a truncated y-axis, or a missing stated denominator) rather than a vague general criticism.

***Q17.*** *Why is 'is the denominator stated' one of the most important questions on this checklist, in your view?*

**A:** Denominator transparency is especially important because a rate or percentage is meaningless, and potentially highly misleading, without knowing the underlying base it was calculated from — a percentage calculated from 4 cases can look identical to one calculated from 4,000 cases despite representing vastly different levels of statistical reliability.

***Q18.*** *Explain why a reviewer should ask whether a plot is presented as a finding or a hypothesis, even if the plot itself is technically accurate.*

**A:** Even a technically accurate plot can be presented in a way that implies more certainty or generality than the underlying analysis supports; asking whether a plot is framed as a finding or a hypothesis checks the FRAMING and claimed strength of evidence, not just the accuracy of what is literally drawn.

***Q19.*** *Propose one additional question you would add to the checklist in Section 5.10 for reviewing a health-related exploratory plot.*

**A:** A reasonable additional question: 'would this chart look meaningfully different with a different, equally defensible choice of binning, grouping, or colour scale?' — checking whether the apparent pattern is robust to reasonable alternative presentation choices, not an artefact of one particular choice.

***Q20.*** *How would you handle giving this kind of critical feedback to a junior colleague or student in a constructive, non-discouraging way?*

**A:** A constructive approach frames the feedback around the shared goal of producing the most defensible, useful analysis possible, asking questions ('what was the denominator for this figure?') rather than making accusatory statements, and explicitly acknowledging the good work already done before raising the specific checklist item that needs attention.

***Q21.*** *Using Section 5.11, explain one advantage an interactive dashboard offers during a collaborative review meeting that a set of static charts cannot.*

**A:** An interactive dashboard allows a meeting participant to immediately filter, drill down, or re-plot the data in response to a question that arises spontaneously during discussion, whereas a fixed set of static charts can only answer the specific questions anticipated in advance by whoever prepared them.

***Q22.*** *Why might a non-technical programme staff member using an interactive dashboard catch an implausible pattern that a data scientist working alone might miss?*

**A:** A non-technical programme staff member with direct, on-the-ground knowledge of a specific facility might notice, while interactively filtering to that facility in real time during the meeting, that a displayed value looks implausible given what they personally know about that facility — a pattern a data scientist without that specific local knowledge might not think to investigate at all.

***Q23.*** *Explain why static charts remain the right choice for a final published report, even if interactive tools were used during exploration.*

**A:** Static charts remain the right choice for a final published report because a report needs to present a fixed, stable, and precisely reviewed set of results that will not change after publication, whereas an interactive dashboard's flexibility (which is its strength during collaborative exploration) is not

compatible with the fixed, citable nature required of a final published document.

***Q24.*** *Propose one risk of relying only on an interactive dashboard for exploration, without ever producing a static, fixed record of what was found.*

**A:** A risk of relying only on an interactive exploration tool is that valuable insights or patterns discovered informally during a live session may not be captured or documented anywhere afterward, meaning the specific finding and the reasoning behind it could be lost or forgotten if not deliberately recorded as a static output at the time.

***Q25.*** *Would you recommend an interactive dashboard for the mini-project in Section 5.9's kind of instrument-translation investigation? Explain your reasoning.*

**A:** Yes — an interactive dashboard could be valuable for the Section 5.9 investigation, allowing the analyst to quickly filter the item-level response distribution by site in real time while discussing findings with the field team, potentially accelerating the discovery of the translation-quality explanation through a live, collaborative back-and-forth rather than a slower cycle of static chart requests and responses.

***Q26.*** *Explain why this chapter treats EDA as a distinct, essential stage rather than something to skip in favour of going straight to formal modelling.*

**A:** EDA is treated as essential because it is frequently the stage at which data quality problems are first detected and at which genuine, testable hypotheses are formed — skipping straight to formal modelling risks building a sophisticated model on top of an unnoticed data problem, or testing a hypothesis that was never grounded in an initial look at the actual data.

***Q27.*** *Summarise the common thread linking the instrument-translation case study (Section 5.9) and the reviewing-someone-else's-plots checklist (Section 5.10).*

**A:** The common thread is that careful, structured examination of a dataset or a set of plots — whether checking whether a translated survey instrument behaves consistently across languages, or systematically reviewing someone else's exploratory plots — can reveal problems (translation artefacts, cherry-picked or misleading visualisations) that a superficial glance would miss.

***Q28.*** *Why does this chapter warn against treating a striking EDA pattern as a finding, even when it looks statistically obvious?*

**A:** A striking EDA pattern is the product of open-ended exploration across many possible relationships, which inflates the chance of finding an apparently compelling but ultimately spurious pattern purely by chance (the multiple comparisons logic introduced formally in Chapter 7); treating it as a finding without further testing risks reporting a false positive with full confidence.

***Q29.*** *Propose one exploratory check you will now perform routinely, having read this chapter, that you did not previously perform.*

**A:** Answers will vary by individual, but should name one specific, concrete exploratory check (e.g., 'always plot the raw distribution of a key outcome variable before summarising it with a single statistic') that the respondent did not previously perform routinely.

***Q30.*** *If you had only one exploratory plot to make before trusting a new dataset, which type would you choose, and why?*

**A:** Answers will vary, but a defensible choice is a simple univariate distribution plot (histogram or boxplot) of the primary outcome variable, since it can reveal skewness, implausible values, or unexpected missingness before any further exploration is worth pursuing.

***Q31.*** *Using Section 5.13, describe a situation from your own work where a dashboard would be more appropriate than a static report, and vice versa.*

**A:** Answers will vary, but a defensible example contrasts an ongoing programme monitoring context (favouring a dashboard, since the M&E team will ask evolving questions over time) against a specific, one-time policy decision context (favouring a report, since a single interpreted answer is needed at one point in time).

***Q32.*** *What is the risk of building an elaborate interactive dashboard for an audience that only needed a two-page memo?*

**A:** The risk is a substantial waste of development effort and technical resources building interactive functionality (filters, drill-downs) that the audience never actually uses, when a much simpler, faster-to-produce two-page memo would have fully met their actual, narrower need.

***Q33.*** *Explain why an audience that will 're-slice' the same data repeatedly is better served by a dashboard than a fixed report.*

**A:** An audience that will re-slice the same data repeatedly benefits from a dashboard because it lets them explore different filters and views on their own schedule without needing to request a new custom report from the analyst every time their question changes.

***Q34.*** *Propose one hybrid approach that could combine some benefits of both a dashboard and a written report for a specific audience.*

**A:** A reasonable hybrid: a short written report presenting the key interpreted findings and recommendation, with an accompanying link to a simpler dashboard for readers who want to explore the underlying data themselves at a more granular level.

***Q35.*** *How should the choice between a dashboard and a report affect the EDA techniques (Sections 5.2-5.11) an analyst emphasises when preparing it?*

**A:** For a dashboard, an analyst should emphasise EDA techniques that generalise well across many possible filter combinations (e.g., small multiples, Section 5.8); for a report, the analyst should emphasise EDA techniques that support a single, carefully chosen, well-interpreted narrative (e.g., one fully annotated key plot, Section 5.12).

## Chapter 6

## Programming Foundations in R and Python for Health Data Science

***Q1.*** *Convert a manual, spreadsheet-based process you currently use for a repeated task (e.g., producing a monthly report for each LGA) into a plan for a scripted, reproducible pipeline.*

**A:** Answers will vary, but a complete plan identifies the repeated manual steps (e.g., filtering by LGA, computing a summary, formatting a table) and describes converting each into a scripted function applied iteratively across all LGAs, with raw data read directly from its source rather than manually copied into the report.

***Q2.*** *Write an R or Python function that takes a facility-month data frame and returns TPR with a flag for implausible values (see Section 4.4), then apply it across multiple states using iteration rather than copy-paste.*

**A:** "'python def flag_tpr(df): df = df.copy() df['tpr'] = df['confirmed_positive'] / df['tests_performed'] df['tpr_flag'] = (df['tpr'] > 1) | (df['tpr'] < 0) return df results = {s: flag_tpr(df[df['state']==s]) for s in df['state'].unique()} "'This applies the same flagging function consistently across every state.

***PYTHON — supporting code***

```
def flag_tpr(df):
    df = df.copy()
    df["tpr"] = df["confirmed_positive"] / df["tests_performed"]
    df["tpr_flag"] = (df["tpr"] > 1) | (df["tpr"] < 0)
    return df

results = {s: flag_tpr(sim[sim["facility_id"].str.startswith(s)]) for s in ["F00",
"F01"]}
```

***Q3.*** *Explain, to a non-programmer colleague, why setting a random seed matters for a machine learning pipeline that will be run again next month on updated data.*

**A:** Setting a random seed ensures that any step involving randomness (splitting data into training/testing sets, bootstrap resampling) produces exactly the same result every time the script is run — this matters because without it, re-running the pipeline next month on updated data could produce a different train/test split purely by chance, making it impossible to tell whether a change in reported performance reflects a real change in the data or just different random luck in the split.

***Q4.*** *In the worked example of Section 6.5, explain what would go wrong (in terms of reproducibility and correctness) if each state's report were instead produced by copy-pasting and manually editing the filter condition 36 times.*

**A:** Copy-pasting and manually editing the filter 36 times risks at least one accidental error (e.g., forgetting to change 'Lagos' to 'Kano' in one copy, so that state's 'report' actually shows Lagos's data), makes it very time-consuming to fix a bug (each of the 36 copies must be found and corrected individually), and leaves no single place where the reporting logic is defined and can be verified for correctness.

***Q5.*** *Using the Python code in Section 6.5, modify 'make_state_report()' to also return the number of months with*

*zero reported tests, and explain why this additional check might be useful for a state-by-state HMIS data quality audit.*

**A:** “‘python def make_state_report(state_name, data): state_data = data[data[‘state’] == state_name] zero_test_months = (state_data.groupby(‘report_month’)[‘tests_performed’].sum() == 0).sum() return {‘mean_tpr’: state_data[‘tpr’].mean(), ‘zero_test_months’: zero_test_months} “‘Adding this check helps flag states/months where reporting may have failed entirely (zero tests recorded), a distinct data-quality issue from low but non-zero testing volume.

***PYTHON — supporting code***

```
def make_state_report(state_name, data):
    state_data = data[data["state"] == state_name]
    monthly_totals = state_data.groupby("report_month")["tests_performed"].sum()
    zero_test_months = int((monthly_totals == 0).sum())
    return {
        "mean_tpr": state_data["tpr"].mean(),
        "total_tests": state_data["tests_performed"].sum(),
        "n_facilities": state_data["facility_id"].nunique(),
        "zero_test_months": zero_test_months,
    }
```

***Q6.*** *Explain, in your own words, the difference between committing all your work in one single Git commit at the end of a project versus committing incrementally after each meaningful step.*

**A:** Committing everything in one enormous final commit provides no intermediate history — if something breaks partway through, there is no earlier working version to revert to, and a reviewer cannot see the logical sequence of changes; incremental commits after each meaningful step create checkpoints that can be individually reverted to, reviewed, and understood.

***Q7.*** *A colleague says version control is 'only for software engineers, not for a statistics thesis.' Write a short response, drawing on the reproducibility arguments made across Chapters 4 and 6.*

**A:** A strong response notes that a statistics thesis, like software, is a codebase that changes over time, is reviewed by others (a supervisor, examiners), and benefits from being able to recover an earlier version after a mistaken edit — version control provides exactly the same auditable history and safety net for an analysis script as it does for software, regardless of the author's job title.

***Q8.*** *Using the systematic debugging approach in Section 6.7, describe the first two steps you would take if a script produced an error you did not immediately understand.*

**A:** The first two steps would be: (1) read the full error message carefully, noting the specific line number and the type of error reported (e.g., a type mismatch, an undefined variable, a dimension mismatch), rather than skimming past it; (2) attempt to isolate the smallest possible piece of code that reproduces the same error, stripping away unrelated parts of the script to focus attention on the actual problem.

***Q9.*** *Explain why isolating the smallest piece of code that reproduces a bug is more efficient than debugging within a full, long script.*

**A:** Isolating the smallest reproducing example is more efficient because debugging within a full script means each test run also executes unrelated code (slowing down each debugging iteration) and requires holding much more context in mind simultaneously, whereas a minimal reproducing example lets the analyst focus entirely on the specific failing logic and iterate much faster.

***Q10.*** *A script runs without any error but produces an obviously wrong summary statistic. Using Section 6.7's fourth step, describe your debugging approach.*

**A:** The appropriate approach is to work backward from the final wrong output, checking each intermediate transformation step's result against what would be expected, printing or inspecting intermediate objects until the exact step where the result first becomes wrong is identified, rather than guessing at the cause from the final output alone.

***Q11.*** *In the Section 6.8 case study, explain why running the refactored pipeline against the original script's results on the SAME original data was an essential verification step, not an optional extra.*

**A:** This verification step is essential because refactoring code — restructuring it for maintainability — is intended to change ONLY the code's structure, not its results; without directly confirming the refactored version reproduces the exact same output on the same input data, there is no way to be confident that the refactoring process itself did not accidentally introduce a subtle change in the analysis's substantive logic.

***PYTHON — supporting code***

```
# A minimal reproducible-seed demonstration, per Section 6.3 and Chapter 3.14
import numpy as np
a = np.random.default_rng(42).normal(size=5)
b = np.random.default_rng(42).normal(size=5)
print("Identical draws:", (a == b).all())
```

***Q12.*** *Describe one risk that would remain even after this verification step, when the refactored pipeline is later run on the new, extended quarter of data.*

**A:** A remaining risk after this verification is that the refactored pipeline, while verified as correct on the ORIGINAL data, could

still contain an edge case or assumption (for example, about date formats or category labels) that only manifests when applied to the new, extended quarter of data, which the original verification step — performed only on the original data — could not have caught.

***Q13.*** *Why does extracting the six copy-pasted plots into a single parameterised function reduce the risk of the specific 'five out of six updated correctly' failure mode described in Section 6.8?*

**A:** A single parameterised function ensures every plot is generated by exactly the same underlying logic every time it is called, so updating that one function's logic automatically and consistently updates all six plots; with six independently copy-pasted and hand-edited blocks, there is no such guarantee, and each block must be individually remembered and correctly updated, creating exactly the risk that one is accidentally missed or updated inconsistently.

***Q14.*** *If you inherited a 400-line ad-hoc script like the one in Section 6.8, would you refactor the whole script at once or incrementally? Justify your choice.*

**A:** A strong response would generally favour incremental refactoring — restructuring and verifying one logical piece (e.g., the data loading step) at a time against the original results, rather than rewriting the whole script at once — because this makes it much easier to isolate the source of any discrepancy that arises during verification, compared to discovering a mismatch only after the entire script has been rewritten.

***Q15.*** *Explain, in your own words, what a 'regression test' means in the data-analysis context described in Section 6.8, as distinct from its meaning in Chapter 8's regression modelling.*

**A:** In Chapter 6's data-analysis context, a 'regression test' means re-running a modified pipeline against a known, previously-verified result to confirm the modification has not

changed that result unexpectedly; in Chapter 8, 'regression' instead refers to the entirely different statistical modelling technique for estimating relationships between variables — the two uses of the word are unrelated beyond sharing the same term.

***Q16.*** *Using Section 6.9, explain the specific reproducibility risk introduced by a notebook's non-linear, cell-by-cell execution model.*

**A:** In a notebook, cells can be run out of the order they appear in the file (for example, re-running an earlier cell after editing a later one), meaning the notebook's displayed output may reflect a sequence of edits and executions that would NOT be reproduced if the notebook were run straight through from top to bottom in a fresh session.

***Q17.*** *Why does 'restart and run all' meaningfully reduce the risk described in Section 6.9, compared to simply saving the notebook after editing?*

**A:** 'Restart and run all' forces the notebook to execute every cell in its written top-to-bottom order in a completely fresh session, directly testing and confirming that the notebook's displayed results are genuinely reproducible from that fixed execution order, rather than being an artefact of whatever ad-hoc sequence of edits and re-runs happened to occur during the original interactive editing session.

***Q18.*** *For which of the tasks in this book (e.g., exploratory analysis, a field-ready SEIR tool) would you personally choose a notebook, and for which a plain script? Justify your choices.*

**A:** A notebook would be well-suited to Chapter 5's exploratory analysis, where interleaving narrative explanation with iterative, interactive plot and summary generation is valuable; a plain script would be better suited to Chapter 18's field-ready SEIR tool, which needs to run reliably, predictably, and repeatedly without any risk of cell-execution-order artefacts.

***Q19.*** *Propose one hybrid workflow that gets some of the benefits of both notebooks and scripts for a single project.*

**A:** A hybrid workflow: conduct initial interactive exploration in a notebook, then, once the analysis approach is settled, refactor the finalised logic into clean, tested functions in a plain script or module (Chapter 6's Section 6.5 pattern) that the notebook itself then imports and calls, combining the notebook's narrative/exploratory strengths with the script's execution-order reliability for the finalised logic.

***Q20.*** *Explain to a student who has only ever used notebooks why a production tool (like the Section 18.11 dashboard) is more often built as a script or application than as a notebook.*

**A:** You would explain that a production tool needs to run correctly and predictably every single time it is invoked, often by someone other than its original author and without any interactive editing at all, and that a notebook's flexible, non-linear editing convenience — its main advantage during exploration — becomes a liability rather than a benefit once the tool needs to be handed off and run reliably by others.

***Q21.*** *Using Section 6.10, write (in words) two test cases you would use to check a function that calculates case fatality ratio (Section 18.5).*

**A:** Test case 1: a dataset with 10 total cases and 5 deaths, all with resolved outcomes, should return a CFR of exactly 50%. Test case 2: a dataset with 10 total cases, 5 deaths, 3 recovered, and 2 unresolved (still hospitalised) outcomes should return a CORRECTED CFR of 5/8 = 62.5% (using only resolved cases), not 5/10 = 50% (which would incorrectly use the naive, uncorrected denominator).

***Q22.*** *Why is testing a function with a deliberately implausible input (e.g., more deaths than total cases) useful, in addition to testing with a normal, expected input?*

**A:** Testing with a deliberately implausible input checks that the function correctly identifies and handles an edge case or error condition it should never silently accept — confirming a function behaves correctly under normal conditions says nothing about whether it also correctly flags or rejects genuinely invalid input, which is an equally important, separate property to verify.

***Q23.*** *Explain how automated tests specifically support the refactoring process described in Section 6.8.*

**A:** Automated tests directly support refactoring by providing an immediate, automatic way to confirm that a refactored version of a function still produces the exact same, previously-verified correct output for the same test inputs — precisely the kind of regression-test verification described qualitatively in Section 6.8, now made systematic and repeatable rather than performed manually and once.

***Q24.*** *Why are automated tests described as 'under-used' in academic health research code, and what barrier do you think most contributes to this?*

**A:** Automated testing is likely under-used in academic health research code because it is often perceived as an extra, non-essential step under time pressure to produce a specific analytic result, and because many academic training programmes in statistics or epidemiology emphasise statistical methodology far more than software engineering practices like testing, leaving many researchers without direct exposure to the practice.

***Q25.*** *Propose one function from any chapter in this book that you would prioritise writing a test for first, and explain your reasoning.*

**A:** A reasonable priority: the TPR-with-plausibility-flag function from Section 4.4, since it is reused repeatedly across

many worked examples in this book and directly affects downstream conclusions in nearly every subsequent chapter that uses malaria diagnostic data, making an error in this specific function unusually consequential if it went undetected.

***Q26.*** *Summarise, in your own words, why this chapter treats programming practice (functions, version control, testing) as inseparable from good statistical analysis.*

**A:** Programming practices such as writing reusable functions, using version control, and testing code are inseparable from good statistical analysis because a statistically correct method implemented in unreliable, unreproducible, or undocumented code cannot actually be trusted or verified by anyone else, including the original analyst at a later date.

***Q27.*** *Explain the common thread linking the refactoring case study (Section 6.8) and the notebook-versus-script discussion (Section 6.9).*

**A:** The common thread is that both refactoring ad hoc code into a maintainable pipeline and choosing between a notebook and a script are really the same underlying question — how should code be organised so that it remains correct, reusable, and comprehensible as a project grows beyond its initial, small-scale exploratory stage.

***Q28.*** *Why does this chapter argue that reproducibility habits benefit even a solo researcher, not just large teams?*

**A:** Reproducibility habits benefit a solo researcher because they are also that researcher's own future self: a script that is undocumented, untested, and not version-controlled will be just as hard for the original author to understand and safely modify six months later as it would be for a new collaborator encountering it for the first time.

***Q29.*** *Propose one habit from this chapter (functions, testing, version control, or documentation) you consider most valuable for your own current work, and explain why.*

**A:** Answers will vary by individual, but should identify one specific habit and justify it based on where the respondent currently perceives the greatest personal risk or inefficiency in their own analytic practice.

***Q30.*** *If you had to teach this chapter's single most important lesson to a new graduate student in one sentence, what would it be?*

**A:** Answers will vary, but a strong one-sentence lesson emphasises that good programming practice is not optional overhead separate from 'the real statistics' — it is a core part of producing trustworthy, reusable scientific work.

***Q31.*** *Explain, using Section 6.12, why 'it works on my machine' is a warning sign rather than a satisfactory outcome.*

**A:** 'It works on my machine' is a warning sign because it means the code's correctness has only been demonstrated under one specific, uncontrolled combination of package versions, and there is no guarantee — and often no way to verify without an environment specification — that the same code will behave identically elsewhere.

***Q32.*** *What specific information does an environment specification (via renv or a Python virtual environment) capture that plain code does not?*

**A:** An environment specification captures the exact versions of every package (and often the language interpreter itself) used in a project, information that plain code alone does not record, since code typically just states a package's name (e.g., 'library(dplyr)') without specifying which version's particular behaviour it was written and tested against.

***Q33.*** *Describe a scenario where a package version mismatch between two collaborators' machines could silently produce different analysis results.*

**A:** A package update between the two analysts' machines could silently change a function's default behaviour or fix/introduce a bug, causing the same code, run on the same data, to produce a different numeric result on each machine without either analyst necessarily realising the discrepancy stems from a version mismatch rather than a data or logic error.

***Q34.*** *Why might capturing an environment specification be considered as important as version-controlling the code itself?*

**A:** An environment specification can be considered just as important as version-controlled code because a perfectly version-controlled script is still not fully reproducible if the packages it depends on are not also pinned to specific versions — both together are needed for genuine, complete reproducibility.

***Q35.*** *If you have never used an environment management tool before, what is the first concrete step you would take to start?*

**A:** A reasonable first concrete step is running 'renv::init()' in an existing R project (or creating a Python virtual environment and generating a 'requirements.txt' file) to capture the current, working set of package versions before making any further changes to the project.

## Chapter 7

## Statistical Inference and Hypothesis Testing in Health Research

***Q1.*** *A study reports p = 0.03 for a mean difference of 0.2 g/dL in haemoglobin between two groups of 5,000 participants each. Discuss statistical versus clinical significance.*

**A:** This is a case where statistical significance (p = 0.03, driven by the very large sample size of 10,000 total participants) does not imply clinical significance: a 0.2 g/dL difference in haemoglobin is unlikely to be clinically meaningful for most

purposes, illustrating that with a large enough sample, even a trivially small true difference will eventually become 'statistically significant.'

***Q2.*** *You test 20 LGAs for a difference in vaccination coverage against a national benchmark. Explain why a Bonferroni or Benjamini-Hochberg correction is appropriate, and compute the Bonferroni-adjusted threshold.*

**A:** A correction is appropriate because running 20 separate tests at alpha = 0.05 each means the overall chance of at least one false positive by chance, if there were truly no differences anywhere, is much higher than 5%. The Bonferroni-adjusted threshold is 0.05/20 = 0.0025 — each individual LGA test must achieve $p < 0.0025$ to be declared significant after correction.

***PYTHON — supporting code***

```
import numpy as np
alpha = 0.05
n_tests = 20
bonferroni_threshold = alpha / n_tests
print(f"Bonferroni-adjusted threshold: {bonferroni_threshold:.4f}")
# Bonferroni-adjusted threshold: 0.0025
```

***Q3.*** *Explain, without formulas, why ignoring cluster design in an NDHS analysis tends to produce overly narrow confidence intervals.*

**A:** Ignoring cluster design treats individuals within the same cluster (village/enumeration area) as if they provide fully independent information, when in fact they tend to be similar to each other (positive intraclass correlation); this effectively means the 'true' amount of independent information is less than the raw sample size suggests, so the naive calculation understates the true standard error and produces overly narrow, falsely precise confidence intervals.

***Q4.*** *Using the worked example in Section 7.6, recompute (by hand or in code) the approximate 95% confidence interval for the birthweight difference if the sample sizes had instead been 60 and 40 (all other figures unchanged), and comment on how the interval width changes.*

**A:** With n1=60, n2=40 (smaller, more unequal samples) and the same means/SDs, the standard error of the difference would be larger than in the original 180/95 example, widening the confidence interval — the interval would likely span roughly zero or close to it in many such recalculations, potentially making the result no longer statistically significant, illustrating how smaller samples reduce precision.

***PYTHON — supporting code***

```
import numpy as np
from scipy import stats

n1, n2 = 60, 40
mean_diff, sd1, sd2 = 140, 410, 450
se = np.sqrt(sd1**2/n1 + sd2**2/n2)
ci_low, ci_high = mean_diff - 1.96*se, mean_diff + 1.96*se
print(f"95% CI: ({ci_low:.0f}, {ci_high:.0f}) g  [wider than the original n=180/95 example]")
```

***Q5.*** *In Section 7.7, explain in your own words why the Benjamini-Hochberg procedure typically flags more LGAs as significant than the Bonferroni correction, for the same set of p-values.*

**A:** Benjamini-Hochberg controls the false discovery rate (the expected proportion of false positives among all rejected hypotheses) rather than the family-wise error rate that Bonferroni controls; BH's sliding, rank-based threshold is less conservative than Bonferroni's single fixed threshold divided by the number of tests, so BH typically retains more true positives (at the cost of tolerating a small, controlled proportion of false positives among the significant results).

***Q6.*** *Using the R code in Section 7.8, explain what a design effect of 3.1 means in practical terms for the effective sample size of a 600-household survey.*

**A:** A design effect of 3.1 means the effective sample size — the sample size that would give the same precision under simple random sampling — is roughly 600/3.1 H" 194 households, not the full nominal 600; in other words, the clustered survey design has 'wasted' a substantial fraction of the nominal sample size's statistical information due to within-cluster similarity.

***Q7.*** *Two RDS estimators applied to the same key-population dataset give HIV prevalence estimates of 14% and 19%. Explain why this discrepancy is not a 'mistake' and how you would report both figures responsibly.*

**A:** This discrepancy reflects genuine methodological uncertainty in RDS estimation — different estimators make different assumptions about the underlying recruitment and network structure, and no single estimator is universally 'correct' for every RDS dataset — so responsible reporting presents the range (14%-19%) explicitly, rather than picking one estimator's figure and presenting it as the single, precise truth.

***Q8.*** *Calculate Cohen's d for the birthweight example in Section 7.6 (mean difference 140g, pooled SD approximately 425g), and classify it as small, medium, or large using the conventional benchmarks in Section 7.9.*

**A:** Cohen's d = mean difference / pooled SD = 140 / 425 H" 0.33, which falls between the conventional 'small' (0.2) and 'medium' (0.5) benchmarks, closer to small-to-medium in magnitude.

### *PYTHON — supporting code*

```
mean_diff, pooled_sd = 140, 425
cohens_d = mean_diff / pooled_sd
print(f"Cohen's d = {cohens_d:.2f}")
# Cohen's d = 0.33 (small-to-medium)
```

***Q9.*** *Explain why an effect size is more comparable across different studies than a p-value alone.*

**A:** An effect size standardises the magnitude of a difference or association in a way that does not depend on sample size, so two studies with very different sample sizes but the same true underlying effect will report similar effect sizes, whereas their p-values could differ enormously purely due to the difference in sample size — making effect sizes more directly comparable across studies.

***Q10.*** *Why might the conventional 'small/medium/large' benchmarks for Cohen's d be inappropriate for judging whether a specific clinical finding is meaningful?*

**A:** Generic Cohen's d benchmarks were derived from broad conventions across the social sciences and do not reflect the specific clinical stakes of a particular outcome; a small Cohen's d could still represent an important clinical difference for a high-stakes, binary outcome (e.g., mortality), while a nominally 'large' Cohen's d might be practically unimportant for an outcome with little clinical consequence — domain judgement, not the generic benchmark alone, should determine practical importance.

***Q11.*** *In the Section 7.10 case study, explain why searching for a subgroup with a larger apparent effect, after seeing the main non-significant result, would have been methodologically inappropriate without pre-specification.*

**A:** Searching for a subgroup with a larger effect after seeing a non-significant main result, without having pre-specified that subgroup analysis in advance, substantially inflates the chance of finding an apparently 'significant' result purely by chance (the multiple comparisons problem from Section 7.3), and reporting only the subgroup that happened to look favourable, without disclosing the broader search, would misrepresent the true strength of evidence to readers.

***Q12.*** *Using the reported 95% CI (-4 to +10 percentage points) in Section 7.10, explain what range of true effects (harmful, none, or beneficial) remains consistent with this study's data.*

**A:** The confidence interval spans a genuine reduction in suppression of up to 4 percentage points (a possible modest harm) at one end, to an improvement of up to 10 percentage points (a moderately large benefit) at the other end — the data are consistent with anything in this fairly wide range, including the possibility of no effect at all.

***Q13.*** *Why is 'a promising trend toward improved suppression' considered an overstatement of the evidence described in Section 7.10?*

**A:** This phrase implies the data provide some meaningful positive signal toward a benefit, when in fact the confidence interval is equally consistent with no effect or even a modest harm; describing a non-significant result with a wide, zero-including confidence interval as a 'promising trend' overstates the actual strength and direction of evidence the study provides.

***Q14.*** *What would a well-justified next step be for the HIV programme, given the non-significant result and wide confidence interval in Section 7.10?*

**A:** A well-justified next step would be to conduct a larger, adequately powered follow-up study (or a meta-analysis combining this study with other similar trials, if any exist) specifically designed to distinguish a plausible true small-to-moderate effect from no effect at all, rather than either abandoning the intervention entirely or prematurely scaling it up nationally based on inconclusive evidence.

***Q15.*** *Explain briefly why a larger sample size, rather than a different statistical test, is generally the correct remedy for the kind of inconclusive result described in Section 7.10.*

**A:** A wider or more precise confidence interval is fundamentally a function of sample size (and, to a lesser extent, of reducing measurement noise); a different statistical test applied to the same underlying sample size and data would not resolve the genuine imprecision in the estimate, since the issue here is a lack of statistical power (insufficient data to detect a modest effect with confidence), not a mismatch between the data and the chosen test.

***Q16.*** *Using Section 7.11, explain why choosing a one-sided test AFTER seeing the data would be methodologically inappropriate, even if the direction chosen matches the data.*

**A:** Choosing a one-sided test after seeing the data would be inappropriate because it retroactively selects the analysis method specifically because it produces a smaller, more favourable p-value for a result already observed, which is a form of the same undisclosed, results-driven analytic choice this book has warned against in the multiple-comparisons and p-hacking discussions of Chapter 7.

***Q17.*** *Give an example of a Nigerian health research question where a one-sided test, pre-specified before data collection, would be scientifically justified.*

**A:** An example: testing whether a newly introduced, biologically well-established antimalarial drug reduces (rather than testing for any difference in either direction in) parasite clearance time compared to the previous first-line drug, when there is strong prior clinical and pharmacological reason to believe an increase in clearance time is essentially implausible.

***Q18.*** *Why does a one-sided test produce a smaller p-value than a two-sided test for the same data and same direction of effect?*

**A:** A one-sided test concentrates the entire 5% rejection region in one tail of the distribution rather than splitting it between both tails (2.5% each, as in a two-sided test), meaning a result

in the pre-specified direction needs to be less extreme to fall within that single, larger one-sided rejection region, producing a smaller p-value for the same underlying data.

***Q19.*** *Explain why two-sided tests are described as the 'more conservative default' for most health research questions.*

**A:** Two-sided tests are the more conservative default because most health research questions do not have strong enough prior certainty to rule out an effect in the 'wrong' direction, and a two-sided test protects against declaring significance for an unexpected result in either direction, which is generally the more scientifically honest and cautious default position to take before seeing the data.

***Q20.*** *If a colleague proposes switching from a two-sided to a one-sided test partway through analysing already-collected data, what would you advise, and why?*

**A:** You would advise against this switch, explaining that changing the test's directionality after the data has already been collected and (implicitly or explicitly) examined undermines the validity of the resulting p-value, since the choice is no longer independent of the observed result — the appropriate test choice must be made and justified before data collection or analysis begins, not adjusted afterward for a more favourable outcome.

***Q21.*** *Using Section 7.12, explain why the HIV counselling study in Section 7.10 might have been designed with too small a sample size for its true effect size.*

**A:** The HIV counselling study may have been designed assuming a larger true effect size than the intervention actually produces, or without a formal power calculation informed by a realistic minimum clinically important effect size at all, resulting in a sample size (240 and 235 participants) too small to reliably distinguish a genuine small-to-moderate effect from no effect, producing the observed wide, inconclusive confidence interval.

***Q22.*** *What does it mean for a study to have '80% power' to detect a specified effect size, in your own words?*

**A:** 80% power means that, if the specified effect size genuinely exists in the population, a study designed with this sample size and this significance threshold would correctly detect (produce a statistically significant result for) that effect in approximately 80% of hypothetical repeated studies — it does not guarantee detection in any single specific study, but describes the long-run probability of successfully detecting a true effect of the specified size.

***Q23.*** *Why is skipping a power calculation described as a 'planning error' rather than simply a matter of limited resources?*

**A:** It is described as a planning error, rather than simply a resource limitation, because a power calculation is a deliberate, foreseeable design step that should be performed BEFORE data collection begins specifically to avoid wasting resources on a study unlikely to produce an informative result — skipping it is a controllable methodological choice, not an unavoidable external constraint.

***Q24.*** *Propose how you would estimate a realistic minimum clinically important effect size before conducting a power calculation for a new study.*

**A:** A reasonable approach is reviewing published literature on similar interventions for a realistic range of previously observed effect sizes, consulting clinical or programmatic experts about what magnitude of effect would actually be considered meaningful enough to justify a change in practice, and using the smaller, more conservative end of that range for the power calculation to avoid underpowering the study.

***Q25.*** *Explain the connection between an underpowered study and the wide, inconclusive confidence interval described in Section 7.10.*

**A:** An underpowered study, by definition, lacks sufficient participants to reliably distinguish a true moderate effect from no effect at all, which directly produces a wide confidence interval (reflecting high uncertainty) around the estimated effect — this is precisely the pattern observed in Section 7.10's example, where the wide interval spanning zero is consistent with the study having been underpowered for the true, possibly moderate, effect size.

***Q26.*** *Summarise, in your own words, why this chapter treats statistical significance and practical importance as two distinct questions that must both be addressed.*

**A:** Statistical significance answers whether an observed effect is likely to be due to chance alone, while practical importance answers whether the effect, if real, is large enough to matter for a real-world decision; this chapter treats them as two distinct questions because a result can be significant but practically trivial, or practically important but not (yet) statistically significant due to limited sample size.

***Q27.*** *Explain the common thread linking the multiple comparisons discussion (Section 7.3) and the power calculation discussion (Section 7.12).*

**A:** The common thread is that a hypothesis test's result depends critically on study design choices made before data collection (how many comparisons will be run, how much statistical power the sample size provides) — decisions that determine how the resulting p-value should be interpreted, rather than the p-value being a self-contained, context-free summary.

***Q28.*** *Why does this chapter emphasise reporting confidence intervals and effect sizes, not just p-values, in every worked example?*

**A:** Reporting confidence intervals and effect sizes alongside p-values in every worked example reflects this chapter's

consistent argument that a p-value alone answers only a narrow question (is this different from the null value) and omits the magnitude and practical relevance information that a confidence interval and effect size directly provide.

***Q29.*** *Propose a research question from your own field where a poorly designed study (e.g., underpowered, or with an inappropriate one-sided test) could produce a misleading conclusion.*

**A:** Answers will vary, but should identify a specific, plausible underpowered or poorly-designed study scenario from the respondent's own field and explain concretely how its design flaw could produce a misleading conclusion.

***Q30.*** *If you had to explain to a policymaker why a 'non-significant' result can still be useful, what would you say?*

**A:** A reasonable response emphasises that a non-significant result, especially from an adequately powered study, provides genuine evidence against a meaningfully large effect, which can itself be valuable and actionable information (for example, ruling out a feared but false safety signal) rather than simply 'no result.'

***Q31.*** *Using Section 7.14, explain in your own words the key conceptual difference between a classical confidence interval and a Bayesian credible interval.*

**A:** A classical confidence interval's correct interpretation concerns the long-run performance of the interval-construction procedure across many repeated hypothetical samples, whereas a Bayesian credible interval can be interpreted directly as a statement about the probability the true parameter lies within a specific range, given the observed data and the assumed prior.

***Q32.*** *Why might a Bayesian approach be particularly well suited to an active outbreak response context?*

**A:** A Bayesian approach suits an active outbreak response because beliefs about key parameters (transmission rate, severity) need to be continuously updated as small amounts of new data arrive in real time, and the Bayesian framework naturally accommodates this sequential updating in a way that repeatedly re-running classical hypothesis tests does not as cleanly support.

***Q33.*** *Describe a scenario where strong, well-established prior knowledge could reasonably inform a Bayesian analysis of a small new dataset.*

**A:** A reasonable example: incorporating a well-established prior estimate of malaria RDT sensitivity from a large prior multi-country meta-analysis when validating the same RDT brand in a new, small local Nigerian study, allowing the small local sample to meaningfully refine, rather than have to independently re-establish from scratch, the sensitivity estimate.

***Q34.*** *What is the common (but technically incorrect, under the classical framework) way people often misinterpret a 95% confidence interval, and how does the Bayesian credible interval avoid this issue?*

**A:** People often (incorrectly, under the classical framework) interpret a 95% confidence interval as 'there is a 95% probability the true value lies in this specific interval' — this interpretation is only strictly valid for a Bayesian credible interval, since the classical framework treats the true parameter as fixed (not random) and the interval itself as the random quantity across repeated sampling.

***Q35.*** *Would you consider a Bayesian or classical approach more appropriate for a large, well-powered national survey with no strong prior information? Explain your reasoning.*

**A:** For a large, well-powered national survey with no strong prior information, a classical approach is often perfectly

adequate and more straightforward to apply and communicate, since the influence of any reasonable prior would be small relative to the large amount of new data, making the choice between frameworks less practically consequential in this specific scenario.

## Chapter 8

## Regression Modelling

## Linear, Logistic, and Beyond

***Q1.*** *A linear regression of birthweight on maternal ANC visits shows a funnel-shaped residual plot. Name the assumption being violated and one remedy.*

**A:** A funnel-shaped (widening) residual plot indicates heteroscedasticity — non-constant error variance, typically increasing with the fitted value. A common remedy is transforming the outcome (e.g., a log transformation) to stabilise the variance, or fitting the model with robust ('sandwich') standard errors that remain valid under heteroscedasticity.

***Q2.*** *Interpret an odds ratio of 1.8 for ITN use (coded 1 = owns net) predicting malaria-negative test result — including the direction of coding.*

**A:** An odds ratio of 1.8 for ITN use (1 = owns net) predicting a malaria-NEGATIVE test result means ITN owners have 1.8 times higher odds of testing negative (i.e., ITN use is associated with protection against malaria) compared to non-owners, holding other variables constant — the direction is protective because the outcome being modelled is a negative (not positive) test result.

***PYTHON — supporting code***

```
import numpy as np
odds_ratio = 1.8
# ORs > 1 for ITN use (1=owns net) predicting a NEGATIVE test is a protective
association
print(f"OR = {odds_ratio}: ITN owners have {odds_ratio:.1f}x the odds of testing
NEGATIVE")
```

***Q3.*** *A Poisson regression of monthly case counts shows residual deviance far exceeding its degrees of freedom. What does this suggest, and what model would you fit instead?*

**A:** Residual deviance far exceeding its degrees of freedom indicates overdispersion — the data show more variability than the Poisson distribution's mean-equals-variance assumption allows, often due to unmeasured heterogeneity or clustering of cases. A negative binomial regression model, which estimates an additional dispersion parameter, would be a more appropriate choice.

***R — supporting code***

```
# R: checking overdispersion (residual deviance / df) and refitting as negative
binomial
library(MASS)
poisson_model <- glm(cases ~ year, family = poisson(link = "log"), data =
surveillance_df)
cat("Deviance/df:", deviance(poisson_model) / df.residual(poisson_model), "\n")
nb_model <- glm.nb(cases ~ year, data = surveillance_df)  # use this if ratio >> 1
```

***Q4.*** *Using the worked example in Section 8.6, explain why saying 'children aged 5-14 are 35% more likely to test positive' is an imprecise restatement of the actual odds ratio finding, and give a more accurate phrasing.*

**A:** Saying '35% more likely to test positive' imprecisely conflates odds with probability/risk; the correct statement is that the ODDS of testing positive are 35% higher, which is not

numerically the same as the probability being 35% higher, especially when the baseline probability is not small. A more accurate phrasing: 'children aged 5-14 have 35% higher odds of a positive test than children under 5, adjusting for season, ITN use, and residence.'

***PYTHON — supporting code***

```
odds_ratio = 1.35
pct_higher_odds = (odds_ratio - 1) * 100
print(f"Odds are {pct_higher_odds:.0f}% higher -- NOT the same as
'{pct_higher_odds:.0f}% more likely'")
```

***Q5.*** *In Section 8.7, explain why the negative binomial model's standard errors are wider than the Poisson model's, even though the point estimates are similar.*

**A:** The negative binomial model's standard errors are wider because it explicitly estimates and accounts for the extra-Poisson variability (overdispersion) present in the data; the Poisson model, by wrongly assuming mean equals variance, understates the true variability and therefore produces artificially narrow (falsely precise) standard errors.

***Q6.*** *A model has an AUC of 0.85 but is poorly calibrated in the highest-risk decile (Section 8.8). Explain, for a specific clinical use case of your choosing, why this combination could still be dangerous in practice.*

**A:** Even with good discrimination (AUC 0.85), a model that under-predicts risk specifically in the highest-risk decile could lead clinicians to under-treat or delay escalation of care for the most severe patients, because the numeric predicted probability shown to them (e.g., 45%) would be reassuringly lower than the true risk (62%) — a dangerous combination for a use case like triage prioritisation for severe malaria.

***Q7.*** *Fit (or describe how you would fit) an interaction between ITN use and residence (urban/rural) in a logistic model of malaria positivity, and explain what a statistically significant interaction term would mean in this context.*

**A:** A statistically significant interaction term between ITN use and residence would mean the protective association of ITN ownership with reduced malaria odds differs meaningfully between urban and rural areas — for example, ITN use might be strongly protective in rural areas (where outdoor/nighttime mosquito exposure is higher) but show a much weaker association in urban areas, and this difference would need to be reported separately for each residence stratum rather than as a single overall ITN effect.

***Q8.*** *Using the checklist in Section 8.9, identify one item you most often see omitted from regression results you have read, and explain why its omission is a problem.*

**A:** Answers will vary by individual observation, but a common and important omission is reporting confidence intervals without also reporting the sample size or the amount of missing data excluded, which makes it impossible for a reader to judge how much of the original dataset the reported estimates actually represent.

***Q9.*** *Explain why stating the reference category for a categorical predictor is essential for correctly interpreting a regression coefficient.*

**A:** Without knowing the reference category, a coefficient's sign and magnitude are uninterpretable — for example, an odds ratio for 'urban residence' only makes sense once the reader knows whether it is being compared against 'rural' or against some other category, and misinterpreting the reference category can lead to a conclusion opposite to what was actually found.

***Q10.*** *A published regression table reports only p-values with no confidence intervals or sample size. Using Section 8.9, list*

*what additional information you would request from the authors.*

**A:** The reader should request: the full model specification (all predictors and the outcome), the sample size and amount of missing data, confidence intervals for the reported effect estimates, and at least one measure of overall model fit or discrimination appropriate to the model type.

***Q11.*** *In the Section 8.10 case study, explain why blood transfusion's very strong statistical association with severe outcome does not mean transfusion is a genuine risk factor for it.*

**A:** Transfusion is administered specifically because a patient is already severely anaemic, which is itself part of, or strongly predictive of, the severe malaria outcome — so transfusion's strong association with severe outcome reflects that it is a downstream marker/consequence of the same underlying severity, not an independent risk factor that causes the severe outcome.

***Q12.*** *Describe the general rule you would apply, when building a predictive model, to decide whether a candidate predictor should be excluded for the reason illustrated in Section 8.10.*

**A:** The general rule: exclude any candidate predictor whose value is only determined, or only becomes clinically meaningful, at or after the point in time the outcome is beginning to unfold — a predictor should only be included if it would genuinely have been available and measurable BEFORE the outcome occurred, at the actual moment a real clinical decision using the model would need to be made.

***Q13.*** *Give one other example (real or plausible) of a health variable that could show a similarly misleading 'reverse causation' pattern in a predictive model.*

**A:** Another plausible example: including 'admitted to intensive care unit' as a predictor of in-hospital mortality — ICU

admission itself is typically a response to a patient already being critically unwell, so it would show a strong association with mortality that reflects reverse causation rather than a genuine, actionable early risk factor.

***Q14.*** *Why is this problem described as being about predictor TIMING rather than about confounding in the sense of Section 8.2?*

**A:** This is about predictor timing specifically because the issue is not that a third, unmeasured variable is driving both transfusion and severe outcome (the classic confounding structure of Section 8.2); rather, transfusion and the severe outcome are causally connected in a straightforward, direct way, just in the opposite direction from what the model implicitly assumes when transfusion is treated as a baseline predictor of the outcome.

***Q15.*** *If a variable's timing relative to the outcome is genuinely ambiguous or unknown, what would you do before including it in a triage risk model?*

**A:** If a variable's timing is genuinely ambiguous, the safest approach is to exclude it from the predictive model until its measurement timing relative to outcome onset can be confirmed with certainty, since including a variable of uncertain timing risks silently reproducing the same reverse-causation problem illustrated in Section 8.10 without the analyst realising it.

***Q16.*** *Using Section 8.11, explain why comparing the raw (unstandardised) coefficients for age (in years) and a 0-27 symptom score directly would be misleading.*

**A:** Comparing the raw coefficients directly would be misleading because a one-year change in age and a one-point change on a 0-27 symptom score represent very different-sized, non-comparable changes in their respective original scales, so a

larger raw coefficient for one predictor does not necessarily mean it has a larger real effect on the outcome than a predictor with a smaller raw coefficient measured on a different scale.

***Q17.*** *Describe, in your own words, how standardising a predictor changes the interpretation of its regression coefficient.*

**A:** Standardising a predictor rescales it so its coefficient is now interpreted as the expected change in the outcome per one-standard-deviation change in that predictor (rather than per one original unit), putting predictors measured in very different original units onto a directly comparable common scale.

***Q18.*** *Why does the chapter caution that a standardised effect size is 'not the same as' practical or clinical importance?*

**A:** This caution exists because a large standardised effect size only reflects the STATISTICAL magnitude of an association relative to that predictor's own observed variability in the specific dataset — it says nothing directly about whether that association is large enough, in absolute or clinical terms, to matter for patient care or programme decisions, which requires separate domain judgment.

***Q19.*** *Propose a scenario where a predictor has a large standardised coefficient but low practical importance, or vice versa.*

**A:** Example: a predictor with very little natural variability across the sample (e.g., nearly everyone has a similar, narrow range of values) could show a large standardised coefficient (since even a small absolute change relative to its narrow standard deviation looks 'large' in standardised terms) despite having limited practical importance because that predictor rarely varies meaningfully across real patients in practice.

***Q20.*** *Would you recommend reporting standardised coefficients, raw coefficients, or both, in a report intended for a mixed clinical and statistical audience? Justify your choice.*

**A:** For a mixed clinical and statistical audience, reporting both is generally the best approach: raw coefficients preserve direct, clinically interpretable units (e.g., 'each additional ANC visit is associated with X grams higher birthweight'), while standardised coefficients allow a fair relative comparison of predictor importance across variables measured on very different original scales.

***Q21.*** *Using Section 8.12, explain why a standard logistic regression on facility-clustered patient data could understate the true uncertainty around a facility-level predictor's effect.*

**A:** A standard logistic regression assumes all observations are independent, but patients treated at the same facility likely share unmeasured facility-level characteristics (staffing, protocols, equipment) that make their outcomes more similar to each other than to patients at a different facility — ignoring this shared structure understates the true uncertainty around any facility-level predictor's effect, producing overly narrow, falsely precise confidence intervals.

***Q22.*** *What does the 'random effect' term in a mixed-effects model represent, in your own words?*

**A:** The random effect term represents facility-to-facility (or patient-to-patient) variation in the baseline outcome rate that is not explained by the model's other, fixed predictors — it allows each facility to have its own estimated baseline tendency toward the outcome, capturing unmeasured facility-level heterogeneity rather than forcing every facility to share exactly the same baseline rate.

***Q23.*** *Explain the connection between mixed-effects models (Section 8.12) and the design-based survey standard errors introduced in Section 7.4.*

**A:** Both mixed-effects models and design-based survey standard errors address the same underlying problem — non-

independent, clustered observations — using different but related statistical machinery: mixed-effects models explicitly model the clustering structure as part of the regression itself, while design-based survey methods (Section 7.4) instead adjust the standard error calculation directly to account for the survey's known clustering design, without necessarily modelling the cluster-level variation explicitly.

***Q24.*** *Propose a dataset from earlier in this book (e.g., the malaria diagnostics or adolescent cohort data) where a mixed-effects model would be more appropriate than a standard regression, and explain why.*

**A:** The malaria diagnostics dataset, with patient-level test results nested within many different facilities across 13 Global Fund-supported states, would be well-suited to a mixed-effects model, since facility-to-facility variation in testing practices, equipment, and patient case-mix likely creates meaningful, unmeasured clustering that a standard (non-mixed) logistic regression would incorrectly treat as independent.

***Q25.*** *Why might ignoring a genuine facility-level clustering structure lead to an overconfident, rather than merely imprecise, regression result?*

**A:** Ignoring genuine facility-level clustering leads to overconfidence, rather than merely imprecision, because the model's standard errors are calculated as if there were more genuinely independent information in the dataset than actually exists — the resulting confidence intervals appear falsely narrow and precise, potentially leading an analyst to declare a finding 'significant' with more confidence than the data actually supports.

***Q26.*** *Summarise, in your own words, why this chapter treats confounding, reverse causation, and clustering as three distinct threats to a regression model's validity.*

**A:** This chapter treats confounding, reverse causation, and clustering as three distinct threats because each requires a different diagnostic and a different remedy: confounding requires adjusting for a third variable, reverse causation requires reconsidering the assumed direction of the association (or the study design itself), and clustering requires accounting for correlated observations (via mixed-effects models or clustered standard errors) rather than treating every row as independent.

***Q27.*** *Explain the common thread linking the reverse-causation case study (Section 8.10) and the mixed-effects models discussion (Section 8.12).*

**A:** The common thread is that a regression model's coefficient can look statistically sound and yet be substantively misleading if the wrong underlying causal story or data structure is assumed — whether because the true causal direction is reversed, or because the model ignores that observations are grouped rather than independent.

***Q28.*** *Why does this chapter insist on checking model assumptions and reporting a full results table, rather than just quoting a headline coefficient?*

**A:** Checking model assumptions and reporting a full results table (following the checklist in Section 8.9) ensures that a reader can independently assess whether the model's conclusions are well-supported, rather than having to take a single headline coefficient on faith without the surrounding context needed to judge its credibility.

***Q29.*** *Propose a regression model from your own field where you would need to check specifically for the reverse-causation pattern illustrated in Section 8.10.*

**A:** Answers will vary, but should identify a specific regression scenario from the respondent's own field where the outcome could plausibly influence the predictor (rather than only the

reverse), and explain why this possibility needs to be explicitly considered or ruled out.

***Q30.*** *If you had to give one piece of advice to someone about to build their first logistic regression model, what would it be?*

**A:** A reasonable single piece of advice: always ask, before interpreting any coefficient, 'could the outcome actually be influencing this predictor, or could an unmeasured third variable be driving both?' — since this question underlies the majority of common regression interpretation errors.

***Q31.*** *Using Section 8.14, explain why a nutrition programme focused on the most severely undernourished children might prefer quantile regression over ordinary linear regression.*

**A:** A nutrition programme focused on the most severely undernourished children cares specifically about the lower tail of the weight-for-age distribution, and quantile regression directly models how a predictor affects that specific tail, whereas ordinary linear regression would only reveal the predictor's effect on the population's average child, which may not reflect the effect on the most vulnerable subgroup at all.

***Q32.*** *Describe a scenario where a predictor's effect could differ substantially between the 10th percentile and the median of an outcome's distribution.*

**A:** A predictor such as household food insecurity could plausibly have a much larger negative association with weight-for-age specifically among the most severely affected children (who may be more sensitive to additional deprivation) than among children near the median of the distribution, who may be more buffered against moderate additional stress.

***Q33.*** *Why would an ordinary mean-focused regression miss the pattern quantile regression is designed to detect?*

**A:** An ordinary mean-focused regression only estimates a single, average effect across the whole distribution, effectively

averaging over any differences in the predictor's effect at different points of the distribution, and would therefore report a blended, potentially misleading single coefficient that obscures the tail-specific pattern quantile regression can reveal.

***Q34.*** *Propose one health outcome, other than weight-for-age, where the tail of the distribution might be more important than the mean.*

**A:** Answers will vary, but a reasonable example is birthweight, where a predictor's effect on the lowest percentiles (very low birthweight infants, at highest risk of adverse outcomes) may be of much greater clinical interest than its effect on the population average birthweight.

***Q35.*** *What practical cost (in terms of interpretability or complexity) comes with fitting quantile regression across several quantiles rather than a single mean regression?*

**A:** Fitting and interpreting quantile regression across several quantiles requires reporting and explaining multiple sets of coefficients (one per quantile) rather than a single set, which is more complex to present and communicate to a non-technical audience than a single mean-based regression result.

## Chapter 9

## Survival Analysis and Time-to-Event Data

***Q1.*** *Explain why treating a participant lost to follow-up at 6 months as if they were event-free for the whole study period would bias a survival analysis.*

**A:** Treating a participant lost to follow-up at 6 months as event-free for the whole study period would incorrectly assume they never experienced the event, when in fact we simply don't know what happened after month 6 — this could bias the survival curve upward (overstating survival) if, in reality, many such participants went on to experience the event shortly after being lost to follow-up.

***Q2.*** *A Cox model shows a hazard ratio of 0.6 for a treatment arm. Interpret this in words, including the direction of the effect on risk.*

**A:** A hazard ratio of 0.6 for the treatment arm means that, at any given point in follow-up, patients in the treatment arm experience the event (e.g., default, death) at 60% of the rate of the comparison arm — i.e., the treatment is associated with a 40% reduction in the instantaneous risk of the event, holding other covariates constant.

***Q3.*** *Why might a Weibull model be preferred over a Cox model when characterising the time course of schistosomiasis reinfection risk after treatment?*

**A:** A Weibull model is preferred when the specific shape of the hazard over time (e.g., low risk immediately post-treatment, rising progressively as re-exposure accumulates) is itself of scientific interest, since it provides an explicit, smooth, interpretable functional form for the hazard and allows extrapolation/simulation of future risk — the Cox model, by contrast, leaves the baseline hazard shape unspecified and is better suited when only the relative (hazard ratio) comparison between groups matters.

***Q4.*** *Using the by-hand Kaplan-Meier calculation in Section 9.5, extend the cohort by one additional patient who is censored at month 6, and recompute S(8) and S(10).*

**A:** Adding a patient censored at month 6 (at risk up to month 6 but not observed to fail there) increases the risk set at months 5-6 but does not change the numerator of events at months 8 or 10; recomputing: at month 5 there are now 6 at risk (5 events among original 5, but with the extra censored patient added to the risk set the multiplier for month 5's step changes slightly) — conceptually, S(8) and S(10) would be very slightly higher than before, because the extra censored patient contributes 'at

risk, no event' information without contributing an additional death/event, gently supporting the survival curve upward at later times relative to a scenario where the same patient's data were simply dropped.

***PYTHON — supporting code***

```
from lifelines import KaplanMeierFitter

durations = [2, 3, 4, 5, 5, 6, 7, 8, 10]        # added a patient censored at month 6
event_observed = [1, 0, 1, 1, 1, 0, 0, 1, 1]
kmf = KaplanMeierFitter().fit(durations, event_observed)
print(kmf.survival_function_)
```

***Q5.*** *In the worked example of Section 9.6, explain what a significant Schoenfeld residual test (p = 0.01) would imply about the appropriateness of reporting a single hazard ratio for HIV co-infection across the whole follow-up period.*

**A:** A significant Schoenfeld residual test (p = 0.01) implies the HIV co-infection effect on treatment default risk is not constant throughout follow-up — for example, it might be much stronger early in treatment (when adherence support has not yet taken effect) and weaker later — meaning a single overall hazard ratio would be a misleading summary, and a time-varying coefficient or a model stratified by follow-up period would be more appropriate.

***Q6.*** *Using the Python code in Section 9.5, modify the 'durations' and 'event_observed' lists to add two more censored patients at months 9 and 11, and describe (without necessarily running the code) how you would expect the tail of the survival curve to change.*

**A:** Adding censored patients at months 9 and 11 would extend the interval over which the survival curve remains at whatever level it reached at month 10 (since the curve only steps down

at event times, not censoring times), and it would slightly widen the confidence band around the tail of the curve at those later times due to the smaller effective risk set, even though the point estimate of S(t) itself would not change at those exact censoring times.

***PYTHON — supporting code***

```
from lifelines import KaplanMeierFitter

durations = [2, 3, 4, 5, 5, 7, 8, 9, 10, 11]      # added censoring at months 9 and 11
event_observed = [1, 0, 1, 1, 1, 0, 1, 0, 1, 0]
kmf = KaplanMeierFitter().fit(durations, event_observed)
kmf.plot_survival_function()
```

***Q7.*** *A colleague proposes simply excluding all censored patients from a survival analysis to 'simplify the data.' Explain, in two to three sentences, what would be lost or biased by doing this.*

**A:** Excluding all censored patients discards real information (that they survived event-free for at least as long as they were observed), which both wastes data and can bias the analysis — for example, if censoring is related to how long a patient survives (informative censoring), excluding them non-randomly changes the composition of the remaining sample and can distort the estimated survival curve in an unpredictable direction.

***PYTHON — supporting code***

```
durations_full = [2, 3, 4, 5, 5, 7, 8, 10]
event_observed_full = [1, 1, 1, 1, 1, 1, 1, 1]     # naive: treat censored as events
durations_correct = durations_full
event_observed_correct = [1, 0, 1, 1, 1, 0, 1, 1]  # correct: respect censoring
# Comparing the two survival functions shows the naive version understates
survival
```

***Q8.*** *Explain, using the malaria treatment failure example in Section 9.7, why treating an unrelated death as ordinary censoring can overstate the true cumulative incidence of treatment failure.*

**A:** Treating an unrelated death as ordinary censoring implicitly assumes the patient would eventually have experienced treatment failure had they lived long enough, which cannot actually be observed and may not be true; this assumption inflates the estimated cumulative incidence of treatment failure because the standard Kaplan-Meier method effectively 'reallocates' the censored patient's risk entirely onto the remaining at-risk population, overstating how many would truly have failed treatment.

***Q9.*** *In what circumstance would a standard Kaplan-Meier analysis (ignoring competing risks) still give an approximately correct answer despite the presence of a competing risk?*

**A:** A standard Kaplan-Meier analysis would give an approximately correct answer for the outcome of interest only when the competing event (e.g., unrelated death) is genuinely rare in the study population, since in that case the amount of overstatement introduced by mishandling the small number of competing events is negligible.

***Q10.*** *Name one other Nigerian health research context (besides malaria treatment failure) where competing risks would need to be considered.*

**A:** Other examples include time to loss of virologic suppression in HIV treatment cohorts (where death from an unrelated cause is a competing risk), or time to relapse after tuberculosis treatment completion (where death from any other cause competes with the relapse event of interest).

***Q11.*** *In the Section 9.8 case study, explain why comparing raw end-of-study suppression percentages across the three*

*treatment eras would be misleading, even before considering any treatment differences.*

**A:** Raw end-of-study suppression percentages would be misleading even before considering treatment differences because patients enrolled in the most recent era have, by construction, had less total follow-up time to achieve suppression before the study's fixed end date than patients from the earliest era — comparing simple percentages ignores this built-in time asymmetry entirely.

***Q12.*** *Why is testing an interaction between treatment era and baseline CD4 count a reasonable thing to explore in this analysis, given what is known clinically about ART regimens?*

**A:** Testing this interaction is reasonable because a newer ART regimen might plausibly show its greatest relative advantage specifically among patients with more advanced immunosuppression (low baseline CD4 count) at enrolment, a clinically meaningful pattern that a model without the interaction term would average away and potentially miss entirely.

***Q13.*** *Explain, in your own words, what it would mean if the proportional hazards check showed the era effect was NOT constant over follow-up time in this study.*

**A:** If the proportional hazards check showed the era effect was not constant over time, it would mean the treatment-era hazard ratio (e.g., a newer era's rate of suppression relative to the earliest) changes meaningfully as follow-up progresses — for example, an early large advantage that narrows over time as other, era-independent service-delivery improvements diffuse system-wide — meaning a single overall hazard ratio would be a misleading, oversimplified summary.

***Q14.*** *Propose one additional covariate (beyond era and CD4 count) you would want included in the Cox model described in Section 9.8, and explain why.*

**A:** A reasonable additional covariate: facility type or location (urban/rural), since service-delivery quality and access can vary meaningfully by facility and could independently influence both which era a patient was more likely to be enrolled in (if facility upgrades were phased in over time) and their probability of achieving suppression.

***Q15.*** *Why does using each patient's actual follow-up time and censoring status give a more accurate era comparison than a simple 'percentage suppressed by end of study' snapshot?*

**A:** Using each patient's actual follow-up time and censoring status correctly credits patients who are still being followed at the end of the study (and have not yet failed to suppress) as 'at risk, no event yet' rather than either wrongly counting them as failures or arbitrarily excluding them, giving a much more accurate picture of the true suppression rate over time than a single end-of-study snapshot could.

***Q16.*** *Using Section 9.9, explain why measuring time from date of diagnosis (rather than date of treatment initiation) could dilute a covariate's true association with time to viral suppression.*

**A:** Measuring from date of diagnosis would combine (1) the variable, often lengthy delay between diagnosis and actually starting treatment, with (2) the separate biological process of achieving suppression once treatment has begun — a covariate that specifically affects only the second process (e.g., a particular drug regimen's potency) could show a diluted association with the combined duration, since much of the total time variability may instead be driven by the first, treatment-initiation-delay process that the covariate has nothing to do with.

***Q17.*** *Propose a research question for which date of diagnosis WOULD be the more appropriate time origin, rather than date of treatment initiation.*

**A:** Date of diagnosis would be the more appropriate time origin for a research question specifically about the full patient journey from diagnosis to suppression, including any delays in linkage to care — for example, evaluating a programme intervention aimed at reducing the overall time from diagnosis to suppression, where the delay-to-initiation itself is part of what is being studied and improved.

***Q18.*** *Why should the choice of time origin be made deliberately and stated explicitly, rather than defaulting to whichever date is most conveniently available?*

**A:** The choice of time origin should be explicit and deliberate because it directly determines what biological or programmatic question the resulting hazard ratios and survival curves actually answer, and an implicit, unstated choice risks the analysis inadvertently answering a different (and possibly less useful, or misleading) question than the one the analyst believes they are answering.

***Q19.*** *Explain, in your own words, how combining two different sub-processes (time to treatment initiation, and time to suppression after initiation) into one duration could obscure a genuine finding.*

**A:** Combining these two sub-processes into a single duration can obscure a genuine finding because a covariate's true, specific effect on one sub-process (say, time to suppression after starting treatment) can be masked, diluted, or even appear reversed in the combined duration if that same covariate has an unrelated or opposite association with the other sub-process (time to treatment initiation) that is bundled into the same measured duration.

***Q20.*** *If a dataset only records date of diagnosis and final outcome, with treatment initiation date missing for many patients, what would you consider doing before proceeding with a survival analysis?*

**A:** With treatment initiation dates missing for many patients, you would first assess whether this missingness itself follows a MAR or MNAR pattern (Chapter 4) related to observed patient characteristics, consider whether multiple imputation of the missing initiation dates is feasible and well-justified, and, if imputation is not defensible, consider restricting the survival analysis to the subsample with complete initiation-date data while explicitly discussing the resulting limitation on generalisability.

***Q21.*** *Using Section 9.10, explain what a frailty term represents in a Cox model fitted to facility-clustered survival data.*

**A:** A frailty term represents an unobserved, shared risk factor affecting all patients within the same facility similarly — for example, an unmeasured facility-level quality-of-care characteristic that makes all of that facility's patients somewhat more or less likely to experience treatment default at any given point in time, beyond what the model's observed, measured covariates can explain.

***Q22.*** *Why would ignoring shared, facility-level unobserved risk factors in a survival analysis risk the same kind of problem as ignoring clustering in ordinary regression (Section 8.12)?*

**A:** Ignoring shared facility-level frailty risks the same problem as ignoring clustering in ordinary regression: patients from the same facility are not truly independent observations, so a standard Cox model that assumes independence will understate the true uncertainty around any facility-level or even individual-level predictor's hazard ratio, producing overly narrow, potentially misleading confidence intervals.

***Q23.*** *What is one alternative to a frailty model, mentioned in Section 9.10, for handling facility-clustered survival data?*

**A:** An alternative mentioned in Section 9.10 is fitting a Cox model stratified by facility, which allows each facility to have

its own distinct baseline hazard function (rather than a shared random effect term), correctly accounting for facility-level heterogeneity in a different but related way to a frailty model.

***Q24.*** *Propose a scenario (beyond ART treatment default) where a frailty model would be an appropriate extension of a standard Cox model.*

**A:** A frailty model would be an appropriate extension for, e.g., a multi-site study of time to relapse after tuberculosis treatment completion, where patients are nested within treatment centres that may differ in unmeasured ways (staff experience, adherence-support practices) affecting relapse risk similarly across all of that centre's patients.

***Q25.*** *Explain, in your own words, why 'a meaningful number of patients nested within a meaningful number of distinct facilities' is the condition that makes a frailty model worth considering.*

**A:** This condition matters because a frailty model's added complexity is only worthwhile when there is a genuinely meaningful clustering structure to account for — with only a handful of patients per facility, or only a couple of facilities involved altogether, there would be too little information to reliably estimate facility-level variation, and a simpler, standard Cox model (or a facility-stratified Cox model) would likely be more practical and equally defensible.

***Q26.*** *Summarise, in your own words, why censoring is described as the defining structural feature that separates survival analysis from ordinary regression.*

**A:** Censoring is the defining structural feature because it means the exact event time for some observations is genuinely unknown (though partially informative — they are known to have survived at least that long), a data structure ordinary regression methods have no built-in mechanism to use

correctly, unlike the survival-specific methods developed throughout this chapter.

***Q27.*** *Explain the common thread linking the competing risks discussion (Section 9.7) and the frailty models discussion (Section 9.10).*

**A:** The common thread is that both competing risks and frailty models address situations where a standard, single-event, independent-observations Cox model's assumptions are violated — respectively, by the presence of a genuinely different competing event type, or by unobserved shared risk factors clustering within groups (such as families or facilities) that make survival times within a group correlated.

***Q28.*** *Why does this chapter emphasise choosing the correct time origin (Section 9.9) as a design decision, not a minor technicality?*

**A:** This chapter treats time-origin choice as a design decision because choosing the wrong origin (for example, calendar time rather than time since a clinically meaningful starting event) can distort the entire survival curve's interpretation, in a way that no amount of correct downstream modelling can fix after the fact.

***Q29.*** *Propose a time-to-event research question from your own field, and identify which extension (competing risks, frailty, or careful time-origin choice) would be most relevant to it.*

**A:** Answers will vary, but should identify a specific time-to-event question from the respondent's own field and correctly identify which of the three extensions (competing risks, frailty, or time-origin choice) is most directly relevant, with a brief justification.

***Q30.*** *If you had to explain to a colleague why 'just excluding the censored patients' is a bad idea, in one sentence, what would you say?*

**A:** A reasonable one-sentence explanation: excluding censored patients discards the real, partial information that they survived event-free for at least as long as they were observed, and can also introduce bias if censoring is related in any way to prognosis.

***Q31.*** *Using Section 9.12, describe a Nigerian clinical pathway (other than tuberculosis treatment) that could be modelled as a multi-state process.*

**A:** A plausible example: an HIV treatment pathway moving from diagnosis, to ART initiation, to viral suppression, to either sustained suppression or virologic failure, with each transition depending on time already spent in the current state and on evolving covariates such as adherence.

***Q32.*** *Why does a multi-state model require more detailed data than a standard single-event survival analysis?*

**A:** A multi-state model requires accurately-timestamped data on every intermediate state transition (not just the final terminal event), which demands considerably more detailed and complete follow-up data collection than a standard survival analysis that only needs to know the time to a single terminal event.

***Q33.*** *Explain how a multi-state model could reveal information that a single time-to-terminal-event analysis would miss.*

**A:** A multi-state model can reveal, for example, that a particular covariate strongly predicts the transition from treatment initiation to early relapse but has little effect on the transition from sustained remission to a later relapse — a distinction a single time-to-terminal-event analysis, which only sees the final outcome, would collapse into one blended effect.

***Q34.*** *What practical data collection challenge would you anticipate in accurately timestamping every intermediate state transition in a real clinical cohort?*

**A:** A practical challenge would be ensuring that every intermediate transition (e.g., the exact date of sputum conversion) is consistently and accurately recorded across all patients and facilities, since incomplete or inconsistent intermediate timestamps would undermine the additional detail a multi-state model is meant to provide.

***Q35.*** *How does the multi-state framework relate conceptually to the competing risks framework introduced earlier in this chapter (Section 9.7)?*

**A:** Multi-state models generalise the competing risks framework: competing risks is the special case of a multi-state model with just one initial state and several possible, mutually exclusive terminal states, while a full multi-state model can include multiple intermediate, non-terminal states between the start and the eventual terminal outcome(s).

## Chapter 10

## Time Series Analysis and Disease Surveillance Forecasting

***Q1.*** *Explain the difference between a trend forecasting task and an outbreak detection task, using a Nigerian disease surveillance example for each.*

**A:** Trend forecasting projects the smooth, expected trajectory of a series over the medium term (e.g., projecting next year's expected leishmaniasis case counts by state for programme budgeting); outbreak detection instead asks, in near-real-time, whether THIS week's count is anomalously high relative to recent history (e.g., an IDSR alert flagging a sudden cholera cluster) — the former supports planning, the latter supports rapid response.

***Q2.*** *A SARIMA model's residuals show significant autocorrelation at lag 12. What does this suggest about the model's adequacy?*

**A:** Significant residual autocorrelation at lag 12 suggests the model has not fully captured the annual seasonal pattern in the data — the seasonal component of the SARIMA specification (or its order) likely needs to be revised, since a well-fitted model's residuals should resemble white noise with no remaining systematic structure at any lag.

***Q3.*** *Why should time series forecasts be evaluated on a chronological hold-out set rather than a random train/test split?*

**A:** Forecasts must be evaluated on a chronological hold-out because time series data have an inherent order and autocorrelation; a random split would allow the model to 'see' future information adjacent to training points (e.g., training on some months from both before and after a test-set month), producing an unrealistically optimistic assessment of how well the model would perform when genuinely forecasting into an unknown future.

***Q4.*** *Using the worked example in Section 10.6, explain why an RMSE of 24 for the SARIMA model compared to 41 for the naive benchmark supports using the more complex model, and describe one scenario in which you might still prefer the simpler naive model despite its higher RMSE.*

**A:** The lower RMSE (24 vs. 41) shows the SARIMA model tracks the true seasonal pattern and trend meaningfully better than simply repeating last year's same-month value, justifying its added complexity for operational forecasting; however, if interpretability or ease of re-fitting each month by non-specialist staff is paramount, and the improvement is only modest, a program might still reasonably prefer the simpler naive benchmark for its transparency and ease of communication.

***R — supporting code***

```
# R: comparing SARIMA to naive at two different RMSE gaps
sarima_rmse <- 24; naive_rmse <- 41
improvement_pct <- (naive_rmse - sarima_rmse) / naive_rmse * 100
cat(sprintf("SARIMA improves on naive by %.0f%%\n", improvement_pct))
```

***Q5.*** *In Section 10.7, explain why the rolling baseline excludes the current week's count itself (using '.shift(1)') when computing the mean and standard deviation for the alert threshold.*

**A:** The current week's own count is excluded from the baseline calculation because including it would mean an unusually high current count directly inflates the very threshold being used to judge whether it is unusual — this would make it harder for a true outbreak to ever exceed its own inflated threshold, defeating the purpose of the alert rule.

***Q6.*** *Modify the logic in Section 10.7 (in words or code) to use a threshold of mean + 3 standard deviations instead of 2, and explain how this would change the trade-off between sensitivity and false alerts.*

**A:** Using mean + 3 SD instead of mean + 2 SD raises the alert threshold, making the rule less sensitive (more likely to miss a real but moderate outbreak) but more specific (less likely to raise a false alarm during ordinary random fluctuation) — an appropriate choice when false alarms carry a high operational cost, and a less appropriate choice when early detection is paramount even at the cost of some false alerts.

***PYTHON — supporting code***

```
import numpy as np, pandas as pd

np.random.seed(0)
weekly_cases = pd.Series(np.random.poisson(10, 104))
weekly_cases[70:74] += np.random.poisson(20, 4)

for n_sd in [2, 3]:
    rolling_mean = weekly_cases.shift(1).rolling(12).mean()
    rolling_sd = weekly_cases.shift(1).rolling(12).std()
    alerts = weekly_cases > (rolling_mean + n_sd * rolling_sd)
    print(f"Threshold mean+{n_sd}SD -> {alerts.sum()} weeks flagged")
```

***Q7.*** *A programme officer asks why a 'simple' rolling-baseline rule is worth using when 'the ARIMA model is more sophisticated.' Write a two-to-three sentence explanation of when the simpler rule is actually the more appropriate tool.*

**A:** A simple rolling-baseline rule is worth using specifically for the near-real-time anomaly-detection task, where speed, transparency, and ease of implementation by front-line surveillance staff (who may not have statistical software or expertise for ARIMA modelling) matter more than long-range forecast accuracy — the ARIMA model, though more sophisticated, is designed to answer a different question (medium-term trend projection) and is not well suited to flagging this week's anomaly in isolation.

***Q8.*** *A facility's monthly report is missing for exactly one month in an otherwise complete 5-year series. Using Section 10.8, describe an appropriate way to handle this gap before fitting a SARIMA model.*

**A:** A single missing month in an otherwise complete 5-year monthly series is a reasonable candidate for interpolation (e.g., linear interpolation between the surrounding months, or substituting the same calendar month's value from the prior year, adjusted for the overall trend) before fitting a SARIMA

model, since a single gap is unlikely to represent a genuine structural change in the underlying process.

***Q9.*** *Explain why silently interpolating over a 6-month reporting gap caused by a known system disruption could mislead a subsequent trend analysis.*

**A:** Silently interpolating over a 6-month gap caused by a known system disruption (e.g., an HMIS platform migration) risks creating a smooth, artificial trend line across a period when the true underlying reporting behaviour genuinely changed, potentially masking a real discontinuity (such as improved or worsened completeness) that a subsequent trend analysis should instead explicitly account for or flag.

***Q10.*** *How would you distinguish, in practice, between a short random reporting gap and a systematic change in reporting completeness?*

**A:** A short random gap typically appears as an isolated missing point with no accompanying change in the pattern or magnitude of surrounding values, whereas a systematic reporting-completeness change is usually accompanied by other evidence — such as a documented system migration date, a sudden shift in the number of facilities reporting, or a persistent step-change in the series' level immediately after the gap — that should prompt investigation rather than routine interpolation.

***Q11.*** *In the Section 10.9 case study, explain why a near-simultaneous rise across many unconnected LGAs is more consistent with a reporting artefact than with a spreading outbreak.*

**A:** A genuinely spreading outbreak would typically show a pattern radiating outward from an index area over time (consistent with person-to-person or vector-borne transmission chains), whereas a near-simultaneous rise across many otherwise-unconnected LGAs is much more consistent with a

shared, system-wide cause affecting all areas at once — such as a new reporting requirement taking effect everywhere on the same date.

***Q12.*** *Why did the analyst check the age and severity profile of newly reported cases, in addition to the timing and geography of the rise?*

**A:** Checking the age and severity profile provided an independent line of evidence: a genuine resurgence driven by declining vaccination coverage would plausibly shift the case profile toward younger, unvaccinated children, whereas an unchanged profile is more consistent with the same underlying population of cases simply being more completely captured by the new reporting system, rather than a new, different pattern of transmission emerging.

***Q13.*** *Explain why 'continued close monitoring for two more months' was a more appropriate conclusion than declaring the matter fully resolved either way.*

**A:** Continued monitoring was more appropriate because the evidence, while leaning toward a reporting artefact, did not completely rule out a modest genuine resurgence occurring simultaneously and being partly masked within the larger reporting-driven increase — declaring the matter fully resolved either way risked either missing a real, ongoing increase in transmission or wrongly continuing to treat an artefact as a genuine outbreak.

***Q14.*** *Propose one additional data source or check that could help distinguish a reporting artefact from a genuine resurgence more quickly than waiting two months.*

**A:** An additional check could compare vaccination coverage trends in the affected LGAs against historical levels, or reach out to a sample of reporting facilities directly to ask whether and when they specifically began complying with the new case-

based reporting requirement, potentially confirming the timing hypothesis more quickly and specifically than waiting for further case data to accumulate.

***Q15.*** *Why can a reporting-system artefact and a genuine small resurgence potentially both be true at the same time, as suggested in Section 10.9?*

**A:** Both can be true simultaneously because an improved reporting system could genuinely be capturing cases that were always occurring but previously went unreported, while a real, separate increase in transmission (for example, from declining vaccination coverage) could also be occurring at the same time — the two explanations are not mutually exclusive, and disentangling their relative contributions may require the kind of multi-pronged investigation described in Section 10.9.

***PYTHON — supporting code***

```
# Illustrating that a resurgence AND a reporting artefact can co-occur
cases_reported_recent = 340
true_baseline_estimate = 310   # after accounting for a suspected reporting catch-up
artefact_contribution = cases_reported_recent - true_baseline_estimate
print(f"Of the apparent rise, roughly {artefact_contribution} cases may reflect reporting catch-up")
```

***Q16.*** *Using Section 10.10, explain why a health system strike lasting three months would violate the stationarity assumption in a way that ordinary differencing cannot fix.*

**A:** A three-month strike would likely cause a sustained drop (or complete absence) in reported cases for that period, followed by a return to a different baseline level or even a temporary surge as a backlog of cases is reported — this is a one-time, discrete-event-driven change in the series' level and reporting behaviour, not a smooth trend or repeating seasonal

pattern, so ordinary differencing (which removes smooth trends) cannot correct for it.

***Q17.*** *Propose how you would model a surveillance series that includes a known, permanent reporting-rate shift partway through, using the indicator-variable approach mentioned in Section 10.10.*

**A:** You would add a binary indicator variable (0 before the shift, 1 from the shift date onward) as an additional covariate in a Poisson/negative binomial GLM trend model (Section 10.3), allowing the model to estimate a distinct level shift at that point rather than forcing a single smooth trend line across the entire series including the disrupted period.

***Q18.*** *Why is visually inspecting a series for a sudden, sustained level shift a useful first diagnostic step before fitting a SARIMA model?*

**A:** Visual inspection is a useful first diagnostic because a genuine structural break typically appears as a sudden, sustained shift in the series' level or variability that persists after the shift (rather than a single unusual point that quickly reverts), and this pattern is often immediately visible to the eye before any formal statistical test is applied, providing a quick initial signal of whether further investigation is warranted.

***Q19.*** *Give an example of an external event (besides a strike or new diagnostic technology) that could cause a structural break in a Nigerian health surveillance series.*

**A:** Other examples: a change in the disease's official case definition partway through the surveillance period, a merger or split of administrative reporting units, or the rollout of a new data collection mobile application replacing paper-based reporting in a phased manner across facilities.

***Q20.*** *If you were unsure whether an apparent change in a series was a structural break or just a strong seasonal peak, how would you investigate further?*

**A:** You would check whether the timing of the apparent change coincides with any known external event or system change (consulting programme records or staff), examine whether the pattern recurs at the same calendar point in prior years (supporting a seasonal explanation) or is a one-time, non-repeating shift (supporting a structural break explanation), and consider consulting the autocorrelation function for a change in its structure before and after the suspected break point.

***Q21.*** *Using Section 10.11, explain why combining forecasts from a SARIMA model, a Poisson GLM, and a naive benchmark tends to be more robust than relying on any single model.*

**A:** Combining forecasts from a SARIMA model, a Poisson GLM, and a naive benchmark tends to be more robust because each model captures a somewhat different aspect of the data-generating process (short-term autocorrelation, covariate-driven trend, and simple seasonal persistence respectively), and their individual errors, arising from different sources, partially cancel out when averaged together, producing a combined forecast less sensitive to any single model's particular weaknesses.

***Q22.*** *Propose a reasonable way to set the weights for each model in an ensemble forecast, based on their recent historical accuracy.*

**A:** A reasonable approach is calculating each model's forecast error (e.g., RMSE) over a recent rolling window of past forecasts, and setting each model's weight in the ensemble inversely proportional to its recent error (so a model that has recently forecast more accurately receives a higher weight), updating these weights periodically as new data and forecast outcomes accumulate.

***Q23.*** *Why is the ensemble forecasting principle described as analogous to the random forest method in Chapter 11?*

**A:** It is analogous to random forests because both methods rely on the same underlying statistical principle — combining multiple individually imperfect predictors (whether decision trees or forecasting models) tends to reduce overall variance and produce a more reliable combined prediction than relying on any single constituent predictor alone.

***Q24.*** *What would you do if one of the three models in an ensemble consistently performed much worse than the other two over several recent forecast periods?*

**A:** If one model consistently underperforms over several recent periods, you would reduce its weight in the ensemble accordingly (potentially toward zero, effectively dropping it), while continuing to monitor whether its performance recovers under different conditions (e.g., during a genuine outbreak versus a stable period) before permanently removing it from consideration.

***Q25.*** *Explain, in your own words, why 'different models make different kinds of errors' is the key idea behind why ensembling helps.*

**A:** The key idea is that no single model is likely to be uniformly best under every condition — one model might excel during stable, seasonal periods while another performs better during an unusual disruption — so averaging across genuinely different modelling approaches captures a broader range of correctly-modelled patterns than committing to any single model's specific assumptions and blind spots.

***Q26.*** *Summarise, in your own words, why this chapter treats trend forecasting and outbreak detection as related but fundamentally different tasks requiring different tools.*

**A:** Trend forecasting and outbreak detection are treated as distinct because they answer different operational questions on different time horizons — trend forecasting projects the

medium-term expected trajectory for planning purposes, while outbreak detection asks in near-real time whether the current observation is anomalous relative to recent history, requiring a different statistical approach and a different tolerance for false alarms versus missed signals.

***Q27.*** *Explain the common thread linking the stationarity discussion (Section 10.10) and the ensemble forecasting discussion (Section 10.11).*

**A:** The common thread is that both stationarity and ensemble forecasting concern the reliability of a forecasting model's underlying assumptions: stationarity asks whether a series' statistical properties are stable enough for a given model class to apply validly, while ensemble forecasting acknowledges that no single model's assumptions are likely to be perfectly correct, and combining several imperfect models' forecasts often outperforms trusting any single one.

***Q28.*** *Why does this chapter insist on benchmarking every forecast against a naive baseline, even when a more sophisticated model is available?*

**A:** Benchmarking against a naive baseline is essential because it is the only way to know whether a sophisticated model's added complexity is actually earning its keep — a model that cannot outperform simply repeating last year's same-period value provides no genuine forecasting value, however impressive its underlying methodology appears.

***Q29.*** *Propose a Nigerian surveillance time series from your own field where distinguishing a genuine resurgence from a reporting artefact (Section 10.9) would be operationally important.*

**A:** Answers will vary, but should identify a specific Nigerian surveillance context (e.g., a seasonal disease with a known reporting-system change) where an apparent resurgence could

plausibly be a reporting artefact rather than a true transmission increase, and explain the operational stakes of getting this distinction wrong.

***Q30.*** *If you had to explain to a non-technical programme manager why a forecasting model needs to be re-validated periodically rather than trusted indefinitely, what would you say?*

**A:** A reasonable explanation: because the relationships a forecasting model learned from past data can change over time (a new intervention, a demographic shift, a reporting system change), a model that performed well historically is not guaranteed to keep performing well, so periodic re-validation against recent, held-out data is necessary to maintain trust in its ongoing forecasts.

***Q31.*** *Using Section 10.13, explain why the most recent one or two weeks of a live surveillance dashboard often show an artificially low case count.*

**A:** The most recent weeks appear artificially low because some cases from those weeks have occurred but have not yet been reported through the surveillance system due to reporting delay — the count for those weeks will typically rise as delayed reports continue to arrive over subsequent weeks, even though no new transmission is occurring during that revision.

***Q32.*** *What information does a nowcasting model need in order to correct for reporting delay?*

**A:** A nowcasting model needs the historical reporting-delay distribution — an empirical record of how many cases from a given past week were still unreported after one week, two weeks, and so on — estimated from the same or a similar surveillance system's past reporting patterns.

***Q33.*** *Describe the risk of a policymaker reacting to an apparent 'recent decline' in a surveillance dashboard without accounting for reporting delay.*

**A:** A policymaker who reacts to an apparent decline without accounting for reporting delay risks prematurely relaxing an intervention or reallocating resources away from an area where transmission has not actually declined, only for the true, delay-corrected case count to be revised upward once the missing reports eventually arrive.

***Q34.*** *Why does a nowcast require a wider uncertainty interval for the most recent period than for periods further in the past?*

**A:** The most recent period requires a wider uncertainty interval because a larger proportion of its true case count is still an inferred extrapolation (based on the historical delay distribution) rather than an actually observed, complete count, whereas periods further in the past have had more time for nearly all their true cases to be reported and are therefore known with much greater certainty.

***Q35.*** *How does nowcasting relate to, but differ from, the forecasting methods introduced earlier in this chapter (Section 10.6)?*

**A:** Nowcasting corrects for incomplete reporting of the recent past (estimating what the true current count already is, given delay), whereas the SARIMA forecasting introduced in Section 10.6 projects genuinely future values that have not yet occurred at all — the two techniques address different sources of 'unknown' data and are often used together in a complete surveillance system.

## Chapter 11

## Machine Learning Fundamentals for Health Prediction

***Q1.*** *Explain, in your own words, why a model with near-perfect accuracy on its training data can be a warning sign rather than a good result.*

**A:** Near-perfect training accuracy is a warning sign because it suggests the model has learned patterns specific to the training

sample (including noise) rather than generalisable relationships — real-world data almost always contains irreducible noise, so a model that fits it almost perfectly is very likely overfit and will perform substantially worse on new, unseen data.

***PYTHON — supporting code***

```
train_accuracy, test_accuracy = 0.99, 0.71
gap = train_accuracy - test_accuracy
print(f"Train-test accuracy gap: {gap:.2f} -- a large gap like this signals
overfitting")
```

***Q2.*** *A dataset has 40 patients per facility across 15 facilities. Explain why a standard (non-grouped) 5-fold cross-validation could overstate model performance.*

**A:** With 40 patients nested within each of 15 facilities, individuals from the same facility likely share unmeasured facility-level characteristics (staffing, equipment, local case mix); a standard 5-fold CV that randomly assigns individual patients to folds would place some patients from the same facility in both the training and validation folds, letting the model partly 'recognise' facility-specific patterns rather than genuinely generalising, thereby overstating true out-of-facility performance.

***Q3.*** *Propose one subgroup analysis you would run before deploying a triage risk model developed on tertiary hospital data to primary health centres.*

**A:** One useful subgroup analysis: compute the model's AUC and calibration separately for tertiary hospital patients versus primary health centre patients, since a model trained predominantly on tertiary data (with more comprehensive lab and vital-sign measurements) may perform substantially worse where such measurements are less consistently available.

***Q4.*** *Using the worked example in Section 11.6, explain in your own words why a 9-predictor LASSO model that performs*

*comparably to a 25-predictor full model is generally preferable in a clinical deployment setting.*

**A:** A parsimonious 9-predictor model is generally preferable because it is easier and cheaper to collect the required data in routine clinical practice, easier to explain and audit for clinical staff, less prone to overfitting on future, slightly different patient populations, and it achieves essentially the same predictive performance as the full 25-predictor model, meaning the extra 16 predictors were adding complexity without meaningfully improving accuracy.

***PYTHON — supporting code***

```
n_predictors_full, n_predictors_lasso = 25, 9
auc_full, auc_lasso = 0.83, 0.82
print(f"LASSO uses {n_predictors_lasso}/{n_predictors_full} predictors for "
    f"{auc_lasso/auc_full:.1%} of the full model's AUC")
```

***Q5.*** *In Section 11.7, explain why the ungrouped cross-validated AUC (0.88) is described as 'optimistic' rather than simply 'better' than the grouped AUC (0.79).*

**A:** The ungrouped AUC is called 'optimistic' rather than simply 'better' because the improvement does not reflect genuinely superior predictive ability for new, unseen facilities — it reflects information leakage (the model partially recognising facility identity via shared patients across folds), so the higher number overstates how well the model would actually perform when deployed at a facility not represented at all during training.

***Q6.*** *Using the R code in Section 11.6 as a guide, describe what would happen to the number of selected variables if you increased the LASSO penalty (lambda) well beyond 'lambda.min'.*

**A:** Increasing the LASSO penalty (lambda) well beyond 'lambda.min' would shrink more coefficients exactly to zero,

resulting in a simpler model with fewer selected variables — potentially fewer than the 9 found at ‘lambda.min‘ — at the cost of some predictive accuracy, since the penalty would be favouring simplicity more aggressively than the value that minimised cross-validated deviance.

***R — supporting code***

```
# R: illustrating how increasing the LASSO penalty shrinks more coefficients to
zero
library(glmnet)
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
for (mult in c(1, 2, 5)) {
  lambda_try <- cv_lasso$lambda.min * mult
  coefs <- coef(cv_lasso, s = lambda_try)
  cat("lambda x", mult, "-> non-zero coefficients:", sum(coefs != 0), "\n")
}
```

***Q7.*** *A hospital administrator asks why your random forest model, which achieved 0.88 AUC in initial testing, performed worse when deployed at a different hospital. Drawing on Sections 11.5 and 11.7, write a short explanation of the likely cause(s).*

**A:** The likely cause is that the model was validated using standard (non-grouped) or otherwise optimistic cross-validation at the original hospital, giving an inflated performance estimate that did not reflect how the model would perform on a genuinely new population; additionally, the new hospital may have a different patient case mix, measurement practices, or facility-level characteristics not represented in the original training data — exactly the deployment risk highlighted in Sections 11.5 and 11.7.

***Q8.*** *Using Section 11.8, propose one engineered feature (beyond age banding) that could improve a severe malaria risk model, and explain the clinical reasoning behind it.*

**A:** One engineered feature: 'time from symptom onset to presentation' relative to the disease's typical progression window, since presenting very late in the disease course is clinically known to be strongly associated with more advanced, higher-risk presentation — a feature the raw age or symptom variables alone would not directly capture.

***Q9.*** *Explain why an interaction feature (e.g., fever duration × danger sign) might capture risk that neither variable alone would capture.*

**A:** An interaction feature can capture risk that emerges specifically from the COMBINATION of two factors — for example, a long fever duration alone might carry modest risk, and a single danger sign alone might carry modest risk, but a patient with both a long fever duration AND a danger sign may be at disproportionately higher risk than either factor would suggest individually, a pattern only a combined (interaction) feature or an interaction term in a model can represent.

***Q10.*** *Give an example of a feature engineering choice that could accidentally leak outcome information into a predictor (creating an unrealistically strong but invalid predictor).*

**A:** An example of outcome leakage: engineering a feature such as 'total length of hospital stay' to predict in-hospital mortality, when length of stay is itself partly determined by the outcome being predicted (patients who die often have a different length-of-stay pattern than survivors) — including such a feature would create an artificially strong but practically useless (and invalid, since it wouldn't be available at the time a real prediction is needed) predictor.

***Q11.*** *In the Section 11.9 case study, explain why an unusually high AUC (0.97) should prompt suspicion rather than celebration.*

**A:** An unusually high AUC (well above what is typically seen for comparable outcomes in the published literature) should

prompt suspicion because it is more likely to indicate an unrealistic advantage from a flawed model specification (such as data leakage) than a genuine breakthrough in predictive accuracy, especially for a widely-studied problem where many previous, competent research teams have not achieved comparable performance.

***Q12.*** *Why is 'mode of delivery' an inappropriate predictor for a model meant to inform risk assessment earlier in antenatal care, even though it is strongly associated with the outcome?*

**A:** Mode of delivery is inappropriate here because it is determined at, or very close to, the time of the actual poor outcome itself (clinicians choose caesarean delivery partly in response to signs the pregnancy is already going poorly), so it could not realistically be known or used at the earlier point in antenatal care when a genuinely useful risk-assessment tool would need to make its prediction.

***Q13.*** *Describe one method (beyond noticing a suspiciously high AUC) you could use to actively check for this kind of data leakage before deployment.*

**A:** One method is to explicitly list every candidate predictor together with the exact time point at which it becomes available relative to the outcome, before any modelling begins, and to exclude (or hold back for a separate research purpose) any predictor that is only available at or after the outcome's occurrence — a predictor-availability audit performed proactively rather than only after an unusually strong result raises suspicion.

***Q14.*** *Would removing the leaked predictor and accepting a lower but more honest AUC (0.79) be the right call even if it meant the model no longer appeared 'state of the art'? Explain your reasoning.*

**A:** Yes — accepting the lower, more honest AUC of 0.79 is the right call, because a model that reports an inflated 0.97 AUC

due to leakage would perform far worse than expected once deployed (since the leaked predictor would not genuinely be available at the real point of prediction), potentially causing serious harm to clinical decision-making and to the credibility of the team and the broader use of predictive models in the institution.

***Q15.*** *Give one other example of a health variable that might be recorded at essentially the same time as, rather than genuinely before, the outcome it is meant to help predict.*

**A:** Another example: including 'number of days hospitalised' as a predictor of in-hospital mortality — patients who die often have a very different (sometimes shorter, sometimes longer, depending on the clinical course) length-of-stay pattern than survivors, and this variable is not meaningfully available at admission, the point at which a real triage prediction would actually need to be made.

***Q16.*** *Using Section 11.10, describe a scenario where two models have identical AUC but one is clearly better calibrated than the other.*

**A:** Scenario: two models both achieve AUC = 0.85 on the same test set; Model A's predicted probabilities closely track observed outcome frequencies at every risk decile, while Model B systematically predicts probabilities close to 0 or 1 (overconfident) even for patients whose true observed outcome rate in that decile is closer to 50% — both models rank patients identically well (same AUC), but Model B's actual predicted probabilities would mislead anyone using them at face value.

***Q17.*** *Why is a calibration curve necessary in addition to, rather than instead of, an AUC score?*

**A:** AUC only measures how well a model RANKS observations from lower to higher risk relative to each other; it says nothing about whether the model's actual predicted probability

VALUES are accurate in an absolute sense, which is precisely what calibration measures — the two metrics answer different, complementary questions and neither can substitute for the other.

***Q18.*** *For which of the two example use cases in this book (a triage risk score shown directly to a clinician, versus a ranking-only prioritisation tool) does calibration matter most? Explain your reasoning.*

**A:** Calibration matters most for the triage risk score shown directly to a clinician, since the clinician will interpret and act on the specific numeric probability displayed (as in the Section 8.8 example); for a ranking-only prioritisation tool (e.g., simply sorting patients from highest to lowest priority for review, without displaying an exact probability), only the ranking (and therefore discrimination/AUC) matters, and calibration is comparatively less critical.

***Q19.*** *Propose one practical fix if a model is found to be well-discriminating but poorly calibrated (systematically overconfident) in its highest-risk decile.*

**A:** A practical fix is recalibration — fitting a simple secondary transformation (such as Platt scaling or isotonic regression) that adjusts the model's raw predicted probabilities to better match observed frequencies, without needing to retrain or change the underlying model's ranking/discrimination behaviour at all.

***Q20.*** *Explain why 'the model correctly ranks patients from lower to higher risk' is not sufficient reassurance if you plan to display the exact predicted probability to a clinician.*

**A:** Correct ranking alone is not sufficient reassurance because a clinician shown a specific probability (e.g., '45% risk') will reasonably interpret and act on that exact number, and if the model is poorly calibrated in that specific risk range, the

clinician is being given a numerically inaccurate risk estimate even though the model would have correctly identified the patient as higher-risk than most others in a purely relative, rank-based sense.

***Q21.*** *Using Section 11.11, explain why a model that always predicts 'no event' could achieve 98% accuracy on a dataset where the outcome occurs in 2% of cases, and why this would be a bad model despite the high accuracy.*

**A:** A model always predicting 'no event' would be correct for 98% of cases simply because 98% of the dataset truly has no event, achieving high accuracy purely through the outcome's rarity — but this model would have zero ability to actually identify a true positive case, making it clinically useless for the specific purpose (early identification of the rare event) it was built for, despite its impressive-sounding accuracy figure.

***Q22.*** *Why is AUC or a precision-recall curve a more appropriate evaluation metric than raw accuracy for an imbalanced classification problem?*

**A:** AUC and precision-recall curves specifically evaluate a model's ability to correctly distinguish positive from negative cases across a range of thresholds, remaining informative even when positive cases are rare, whereas raw accuracy can be dominated entirely by how well a model predicts the majority class, providing essentially no useful information about performance on the minority class of actual clinical interest.

***Q23.*** *Describe the difference between oversampling and undersampling as remedies for class imbalance.*

**A:** Oversampling artificially increases the number of minority-class (rare-outcome) examples in the training data (e.g., by duplicating or synthetically generating additional minority-class records), while undersampling instead reduces the number of majority-class examples used in training — both aim to give

the model more balanced exposure to both classes during training, though each carries different trade-offs (oversampling risks overfitting to duplicated examples; undersampling discards potentially useful majority-class information).

***Q24.*** *How is cost-sensitive learning, mentioned in Section 11.11, connected to the classification threshold discussion in Section 15.10?*

**A:** Cost-sensitive learning explicitly assigns a higher penalty to the model for missing a rare positive case (a false negative) than for a false alarm (a false positive) during the training process itself, directly encoding the same asymmetric cost reasoning discussed for setting a classification threshold in Section 15.10, but applied during model TRAINING rather than only when choosing a threshold after training is complete.

***Q25.*** *Propose a Nigerian health prediction task (besides those mentioned in Section 11.11) where class imbalance would be a serious concern, and explain why.*

**A:** A plausible example: predicting maternal near-miss events (severe, life-threatening but non-fatal obstetric complications) from routine antenatal and delivery data — these events are, fortunately, relatively rare, making class imbalance a serious concern for any model attempting to flag at-risk pregnancies for closer monitoring.

***Q26.*** *Summarise, in your own words, why this chapter treats the bias-variance tradeoff as the single most important conceptual idea in machine learning.*

**A:** The bias-variance tradeoff is treated as central because nearly every other technique in the chapter — regularisation, cross-validation, ensemble methods — is, at its core, a strategy for managing this single underlying tension between a model too simple to capture the true pattern and one too complex to generalise beyond its training data.

***Q27.*** *Explain the common thread linking the class imbalance discussion (Section 11.11) and the calibration curves discussion (Section 11.10).*

**A:** The common thread is that both class imbalance and calibration concern situations where a model's headline performance metric (such as overall accuracy or AUC) can look acceptable while masking a serious practical problem — a model can achieve high accuracy on an imbalanced dataset simply by favouring the majority class, and can achieve high AUC while still being poorly calibrated for the specific risk level that matters most operationally.

***Q28.*** *Why does this chapter warn that a model which 'worked too well' (Section 11.9) should prompt suspicion rather than celebration?*

**A:** A model that 'worked too well' should prompt suspicion because near-perfect performance, especially unusually higher than comparable published results, is far more often a sign of information leakage, overfitting, or a data problem (such as the grouping/clustering issue discussed in Section 11.7) than a sign of a genuinely superior method.

***Q29.*** *Propose a prediction task from your own field where class imbalance (Section 11.11) would likely be a serious concern, and describe one way you would address it.*

**A:** Answers will vary, but should identify a specific prediction task with a rare outcome (e.g., a severe adverse event, a rare disease) and describe an appropriate remedy such as resampling, class-weighted loss functions, or reporting precision-recall rather than accuracy alone.

***Q30.*** *If you had to give one warning to a colleague excited about a new machine learning model's high accuracy, what would it be?*

**A:** A reasonable warning: ask what cross-validation strategy was used and whether it respects any natural grouping in the data, since a model's headline accuracy figure alone says nothing about whether that figure will hold up on genuinely new, unseen data.

***Q31.*** *Using Section 11.13, explain why it is generally impossible to satisfy demographic parity, equalised odds, and predictive parity simultaneously.*

**A:** These three criteria generally cannot all be satisfied simultaneously (except in special cases, such as when the outcome's true prevalence is identical across subgroups) because they impose different, mathematically distinct constraints on the relationship between predictions, true outcomes, and subgroup membership, and satisfying one exactly typically requires violating at least one of the others when true outcome rates differ across subgroups, as they very often do in real health data.

***Q32.*** *Describe a clinical scenario where equalised odds might be the most ethically appropriate fairness criterion to prioritise.*

**A:** A reasonable scenario: a screening test used to decide who receives further diagnostic workup, where missing a true case (a false negative) in either subgroup carries a serious health cost, making it important that the true positive rate (sensitivity) be equal across subgroups (part of equalised odds), so that neither subgroup is disproportionately failed by missed diagnoses.

***Q33.*** *Describe a different clinical scenario where predictive parity might be more appropriate instead.*

**A:** A reasonable scenario: a resource-constrained referral system where a positive prediction directly triggers a costly, limited-availability intervention, making it important that a

positive prediction mean approximately the same thing (the same probability of true need) across subgroups (predictive parity), so that the intervention is allocated with consistent reliability regardless of which subgroup a patient belongs to.

***Q34.*** *Why is it important for a health data scientist to make an explicit, justified choice among fairness criteria rather than assuming one universal definition of fairness exists?*

**A:** An explicit, justified choice is important because different fairness criteria protect against different kinds of harm, and silently defaulting to whichever criterion a software package happens to report (often subgroup AUC alone) may not reflect the ethical priority most relevant to a specific clinical application — this decision has real consequences and should be made deliberately and documented.

***Q35.*** *How does the concept of demographic parity relate to, but differ from, the subgroup AUC and calibration checks introduced in Chapter 16?*

**A:** Demographic parity requires an equal overall positive prediction rate across subgroups regardless of true underlying need, which is a different (and sometimes conflicting) standard from checking whether a model's AUC and calibration are similarly accurate across subgroups, as in Chapter 16's audit — a model could have equal AUC across subgroups while still violating demographic parity, or vice versa.

## Chapter 12

## Unsupervised Learning

### Clustering and Dimensionality Reduction

***Q1.*** *Explain why variables should be standardised before k-means clustering, using an example where one variable is measured in the hundreds (e.g., CD4 count) and another is a proportion between 0 and 1.*

**A:** Standardisation is needed because k-means uses distances between points to assign cluster membership, and a variable measured in the hundreds (e.g., CD4 count, ranging from near 0 to 1000+) would dominate the distance calculation compared to a proportion variable ranging only from 0 to 1, effectively causing the clustering to be driven almost entirely by CD4 count regardless of the other variable's true importance — standardising both to comparable scales (e.g., mean 0, SD 1) ensures each variable contributes proportionately.

***PYTHON — supporting code***

```
import numpy as np
cd4 = np.array([250, 800, 120, 950])          # raw scale: hundreds
proportion = np.array([0.2, 0.8, 0.1, 0.95])   # raw scale: 0-1
# Without scaling, cd4's large range would dominate any Euclidean distance
calculation
from sklearn.preprocessing import StandardScaler
scaled = StandardScaler().fit_transform(np.column_stack([cd4, proportion]))
print(scaled)
```

***Q2.*** *You cluster health facilities into 4 groups by testing profile. Describe two ways you would check whether this clustering is meaningful, beyond the silhouette coefficient.*

**A:** Beyond the silhouette coefficient, one should check whether the resulting clusters correspond to a substantively meaningful and actionable distinction recognisable to domain experts familiar with the facilities (e.g., do the four clusters correspond to genuinely different service-delivery profiles that a quality-improvement team would recognise?), and check the stability of the cluster assignments under bootstrap resampling or small perturbations of the input data.

***Q3.*** *Explain, conceptually, how PCA could be used to construct a single household wealth index from a 15-item asset ownership questionnaire.*

**A:** Running PCA on the 15 binary asset items would identify the first principal component (the linear combination of items explaining the most shared variance), which typically loads positively on nearly all asset items and can therefore be interpreted as a general wealth dimension; each household's score on this component becomes a single continuous wealth index, which can then be divided into quintiles for reporting, exactly as the DHS wealth index is constructed.

***Q4.*** *Using the worked example in Section 12.6, explain why converging evidence from both the elbow method and the silhouette score is more persuasive than either method alone for choosing k=4.*

**A:** Converging evidence from two independent diagnostics (a sharp elbow AND a peak silhouette score both at k=4) is more persuasive than either alone because each method has different weaknesses — the elbow method can be subjective to interpret, and silhouette score alone can sometimes favour a trivial or overly simple partition; when both independently point to the same k, this substantially strengthens confidence that k=4 is a genuinely appropriate choice rather than an artefact of one particular diagnostic.

### *PYTHON — supporting code*

```
elbow_k, silhouette_k = 4, 4
print("Elbow method and silhouette score agree:" , elbow_k == silhouette_k)
# Convergent evidence from two independent diagnostics is stronger than either
alone
```

***Q5.*** *In Section 12.7, explain what it means for the first principal component to 'load positively on nearly all items,' and why this supports interpreting it as a general wealth dimension rather than, say, a rural/urban distinction.*

**A:** Loading positively on nearly all items means that households scoring high on this component tend to own most or all of the

assets, and households scoring low tend to own few or none — a pattern consistent with a single underlying 'more resources overall' (wealth) dimension, rather than, say, a rural/urban distinction, which would instead be expected to show some items loading positively and others (e.g., items more common in one setting than the other) loading negatively or near zero.

***Q6.*** *Using the R code in Section 12.7 as a guide, describe what would change in your interpretation if the second principal component (rather than the first) showed strong positive loadings on rural-associated items and negative loadings on urban-associated items.*

**A:** If the second component instead showed strong positive loadings on rural-associated items and negative loadings on urban-associated items, this would suggest the second component captures a distinct rural-urban lifestyle dimension (independent of overall wealth), meaning some households could score similarly on the first (wealth) component but differ substantially on this second, urbanicity-related dimension — a genuinely different and additional piece of information not captured by the wealth index alone.

***PYTHON — supporting code***

```
import numpy as np
# Comparing silhouette scores across k to confirm the elbow choice of k=4
k_values = [2, 3, 4, 5, 6]
wcss =       [820, 540, 310, 295, 280]
silhouette =  [0.31, 0.44, 0.52, 0.38, 0.31]
best_k = k_values[int(np.argmax(silhouette))]
print(f"Best k by silhouette score: {best_k}")
```

***Q7.*** *A colleague clusters patients into 6 groups because 'more clusters gives more detail.' Using the logic of Sections 12.5-12.6, explain why more clusters is not automatically better.*

**A:** More clusters is not automatically better because increasing k will always reduce within-cluster sum of squares (fitting the

data more closely) even when the additional splits are not substantively meaningful — beyond the point where a genuine, interpretable subgroup structure exists (as identified via the elbow and silhouette diagnostics), further splitting tends to divide genuinely similar observations into arbitrary, unstable, and hard-to-interpret subgroups that add complexity without adding real insight.

***Q8.*** *Using Section 12.8, explain why DBSCAN might be more appropriate than k-means for clustering the geographic locations of outbreak cases.*

**A:** DBSCAN would be more appropriate for outbreak case locations because it can find irregularly-shaped clusters (e.g., cases distributed along a river or a road, rather than in a neat circular pattern) and does not require pre-specifying the number of clusters, unlike k-means, which assumes roughly spherical, equally-sized clusters and would force every case into some cluster regardless of true spatial structure.

***Q9.*** *What does it mean for DBSCAN to label some points as 'noise' rather than assigning them to a cluster, and why might this be a desirable property for outbreak case mapping?*

**A:** Labelling isolated points as 'noise' rather than forcing them into the nearest cluster means DBSCAN correctly represents cases that are genuinely spatially separate from any dense outbreak cluster (for example, an unrelated sporadic case or a case linked to travel rather than local transmission) rather than artificially attributing them to a nearby cluster they don't truly belong to — this distinction is directly useful for identifying a genuine outbreak hotspot separately from background sporadic cases.

***Q10.*** *Describe one limitation of DBSCAN compared to k-means (for example, in terms of parameter selection or its behaviour with clusters of very different densities).*

**A:** A limitation of DBSCAN is that its two key parameters (the neighbourhood radius and the minimum number of points required to form a dense region) can be difficult to choose appropriately, and DBSCAN can struggle when a dataset contains clusters of very different densities, since a single radius/minimum-points setting appropriate for a dense cluster may fail to detect a genuinely real but sparser cluster elsewhere in the same dataset.

***Q11.*** *In the Section 12.9 case study, explain why the mismatch between the clustering result and the official facility tier system was treated as a finding rather than a failure.*

**A:** The mismatch was treated as a finding, not a failure, because the goal of the clustering exercise was to discover genuine empirical structure in the facilities' actual operational characteristics, not to reproduce a pre-existing administrative label — and a clustering result that diverges informatively from expectation, when properly investigated, often reveals something substantively true and useful rather than indicating the method has gone wrong.

***Q12.*** *What practical actions might a state health administration take in response to discovering that some 'secondary' facilities empirically resemble 'primary' facilities?*

**A:** Practical actions could include reviewing and potentially updating the official tier classification for the mismatched facilities, directing targeted infrastructure or staffing investment to nominally 'secondary' facilities that empirically function more like primary facilities, or adjusting referral protocols to reflect facilities' actual current capacity rather than their outdated administrative designation.

***Q13.*** *Why would it have been a mistake for the research team to force the clustering to match the expected tier categories rather than reporting the empirical result?*

**A:** Forcing the clustering to match the expected tier categories (for example, by constraining the number of clusters or manually overriding cluster assignments) would have suppressed exactly the useful signal the analysis was capable of revealing, substituting the team's prior expectation for the empirical evidence the data actually provided.

***Q14.*** *Propose one additional facility-level indicator that could be added to strengthen this clustering analysis, and explain what it might reveal.*

**A:** An additional useful indicator could be a measure of recent patient outcomes or clinical quality (e.g., rate of successful emergency obstetric referrals), which would help distinguish facilities that merely have similar infrastructure metrics from those that also perform similarly in actual clinical practice, adding a further, outcome-oriented dimension to the clustering.

***Q15.*** *How would you communicate this finding sensitively to facility staff, given that it could be seen as a critique of their facility's official classification?*

**A:** This finding should be communicated collaboratively and non-punitively, framing it as an opportunity to better align resources and support with each facility's actual current operating reality, rather than as a criticism of facility staff, who typically have little control over their facility's official administrative classification or the infrastructure investment decisions that shape it.

***Q16.*** *Using Section 12.10, explain what a low or negative individual silhouette score indicates about a specific observation's cluster assignment.*

**A:** A low or negative individual silhouette score indicates that a specific observation is not well-matched to its assigned cluster — it may be roughly equidistant from two clusters, or even closer, on average, to a neighbouring cluster's members

than to its own cluster's members, making its cluster assignment genuinely ambiguous or borderline rather than clear-cut.

***Q17.*** *Why might a facility with a low individual silhouette score be worth individual, qualitative review rather than automatic reassignment to another cluster?*

**A:** A facility with a low individual silhouette score sits at the boundary between two archetypes and may not fit either category's typical profile or intervention needs cleanly, meaning a purely mechanical assignment to whichever cluster it happens to be marginally closer to could misdirect quality-improvement resources; individual, qualitative review allows staff with direct facility knowledge to make a more informed judgment about that specific borderline case.

***Q18.*** *Propose how you would use individual silhouette scores to prioritise which facilities a quality-improvement team should visit first.*

**A:** A quality-improvement team could rank facilities by ascending individual silhouette score and prioritise site visits to the lowest-scoring (most ambiguous) facilities first, since these are the cases where the automated clustering assignment is least confident and where on-the-ground context is most likely to add genuine value beyond what the clustering algorithm alone can determine.

***Q19.*** *Explain the difference between the average silhouette score (Section 12.6) and the individual silhouette score (Section 12.10), and what each is used for.*

**A:** The average silhouette score (Section 12.6) summarises the OVERALL quality of a chosen clustering solution across all observations, used to compare and choose between different values of k; the individual silhouette score (Section 12.10) instead evaluates each SPECIFIC observation's fit within its

assigned cluster, used to identify particular borderline cases worth closer review, a distinct and complementary diagnostic purpose.

***Q20.*** *If many observations in a cluster solution have low individual silhouette scores, what might this suggest about the overall choice of k?*

**A:** If many observations have low individual silhouette scores, this suggests the overall chosen value of k may not be well-suited to the data's true underlying structure — either too many or too few clusters have been specified relative to how naturally separated the data actually is — and revisiting the choice of k (Section 12.6) would be a reasonable next step.

***Q21.*** *Using Section 12.11, explain why a clustering result should be treated as an input to domain expert judgment rather than a final, self-evidently meaningful answer.*

**A:** A clustering result should be treated as an input to domain expert judgment because the algorithm itself has no understanding of the real-world, substantive meaning of the groupings it produces — it can only organise data based on the numeric similarity of the variables provided, and translating that numeric grouping into an actionable, correctly-interpreted real-world category requires human expertise the algorithm does not possess.

***Q22.*** *What risk does a data scientist take by presenting a clustering result as a final answer without a collaborative interpretation step?*

**A:** A data scientist who presents a clustering result as a final answer risks either overstating the algorithm's autonomous insight (implying it has 'discovered' a meaningful category when it may simply reflect an artefact of the specific variables chosen) or missing the opportunity for a domain expert to catch a spurious or misleading grouping before it influences a real resource-allocation or policy decision.

***Q23.*** *Propose how you would structure a meeting with facility managers to interpret a new clustering result together, drawing on Section 12.9's case study.*

**A:** A reasonable structure: present the clustering result (e.g., the four facility archetypes and which facilities fall into each) without pre-labelling the clusters, ask facility managers to describe what they notice is similar or different about the facilities within a given cluster based on their own direct knowledge, and jointly develop a substantive, meaningful label and interpretation for each cluster only after this discussion.

***Q24.*** *Why might even a methodologically flawless clustering analysis (correctly validated with elbow and silhouette diagnostics) still need substantive human interpretation?*

**A:** Even a methodologically flawless clustering analysis only guarantees that the chosen groupings are STATISTICALLY well-separated and internally consistent given the specific variables used — it says nothing about whether those groupings correspond to a meaningful, actionable, or previously-unrecognised real-world distinction, which is a substantive judgment that statistical validation alone cannot make.

***Q25.*** *Explain, using your own words, the difference between an algorithm 'surfacing a pattern' and an algorithm 'producing a final answer.'*

**A:** An algorithm 'surfacing a pattern' means it organises and highlights structure in the data for further human consideration; an algorithm 'producing a final answer' would imply the pattern it finds is automatically and completely interpretable and actionable without further review — this book consistently argues clustering algorithms should be understood in the former sense, not the latter.

***Q26.*** *Summarise, in your own words, why this chapter insists that clustering results must be validated against domain expertise, not statistical criteria alone.*

**A:** This chapter insists on domain-expert validation because internal statistical criteria (silhouette score, within-cluster sum of squares) can only assess whether clusters are mathematically well-separated in the chosen feature space — they cannot assess whether the resulting groups correspond to a real, meaningful, and actionable distinction in the actual population being studied.

***Q27.*** *Explain the common thread linking the individual silhouette scores discussion (Section 12.10) and the extended case study of clusters telling an unexpected story (Section 12.9).*

**A:** The common thread is that a clustering result's apparent statistical quality can diverge sharply from its substantive meaningfulness — individual silhouette scores can reveal that some observations are poorly matched to their assigned cluster even when the average score looks acceptable, and a clustering result can technically satisfy every statistical criterion while still telling an unexpected or even implausible substantive story.

***Q28.*** *Why does this chapter describe cluster analysis as an input to, rather than a replacement for, domain expertise (Section 12.11)?*

**A:** Cluster analysis is described as an input to domain expertise, not a replacement for it, because the algorithm can only group observations based on the specific numeric features provided to it — it has no capacity to judge whether the resulting groups make clinical, programmatic, or biological sense, a judgement only a domain expert familiar with the underlying population can make.

***Q29.*** *Propose a clustering task from your own field where you would want to examine individual silhouette scores, not just the average, before trusting the result.*

**A:** Answers will vary, but should identify a specific clustering application (e.g., patient subtyping, facility profiling) where a

small subset of individual observations with low silhouette scores could represent a genuinely important edge case worth separate investigation, rather than simply averaging over them.

***Q30.*** *If a colleague presented a clustering result and asked whether it was 'correct,' how would you respond, drawing on this chapter's arguments?*

**A:** A reasonable response explains that a clustering result is never simply 'correct' or 'incorrect' in an absolute sense — it should be evaluated on whether it is statistically defensible (via the internal criteria of Sections 12.5-12.6) and substantively meaningful and stable (via domain review and the bootstrap check of Section 12.12), with both dimensions needed before the result is trusted.

***Q31.*** *Using Section 12.13, explain the difference between a scree plot's 'elbow' and the Kaiser criterion for choosing how many principal components to retain.*

**A:** The scree plot's elbow is a purely visual, informal judgement about where the marginal variance explained by additional components drops off sharply, while the Kaiser criterion is a specific, numeric rule (retain components with eigenvalue greater than 1) that can be applied mechanically without visual judgement, though the two often (but not always) suggest a similar number of components.

***Q32.*** *Why are both of these methods described as heuristics rather than definitive statistical tests?*

**A:** Both are described as heuristics because neither is derived from a formal statistical hypothesis test with a rigorously justified error rate — they are practical rules of thumb that have proven useful in practice but can occasionally give different or debatable answers, especially in borderline cases where the variance-explained curve does not show a sharp, unambiguous elbow.

***Q33.*** *Explain why substantive interpretability should also inform how many components are retained, not just the numeric criteria alone.*

**A:** Substantive interpretability should inform the decision because a component that is difficult to interpret in terms of any coherent underlying concept (unlike, for example, a clear wealth dimension in Section 12.7) provides little practical value even if it explains a statistically non-trivial amount of variance, and including uninterpretable components can complicate rather than simplify subsequent analysis and communication.

***Q34.*** *Describe a scenario where the scree plot and the Kaiser criterion might suggest a different number of components to retain, and how you would decide between them.*

**A:** If the scree plot suggests retaining 3 components (a clear elbow after the third) but the Kaiser criterion (eigenvalue > 1) suggests retaining 5, a reasonable approach is to examine whether the 4th and 5th components are substantively interpretable and meaningfully improve any downstream analysis; if not, the more parsimonious 3-component solution from the scree plot is often preferred.

***Q35.*** *How does the logic of choosing the number of principal components relate to the logic of choosing k in k-means clustering (Section 12.6)?*

**A:** Both problems share the same underlying logic: balancing a purely mathematical/statistical criterion (variance explained, or within-cluster sum of squares) against the practical and substantive value of simplicity and interpretability, using multiple imperfect heuristics together rather than relying on any single automatic rule.

## Chapter 13

## Mathematical Modelling of Infectious Diseases

***Q1.*** *Explain, in terms of R0, what it means for an outbreak to be 'controlled' by an intervention that reduces the effective transmission rate.*

**A:** An intervention that reduces the effective transmission rate lowers R0 (since R0 is proportional to the transmission rate divided by the recovery rate); if the intervention is strong enough to push R0 below 1, each infectious case produces, on average, fewer than one secondary case, meaning the outbreak will decline over time and eventually die out rather than continuing to grow — this is the formal definition of an outbreak being 'controlled.'

***Q2.*** *Why does modelling malaria transmission require a vector compartment, unlike modelling measles transmission?*

**A:** Malaria transmission requires a mosquito vector as an intermediate host between infected and susceptible humans — an infected person does not directly transmit malaria to another person on contact — so the model must explicitly track the mosquito population's infection status and biting behaviour, whereas measles transmits directly from person to person through respiratory droplets and requires only a human-to-human compartmental structure.

***Q3.*** *Contrast what a fitted SEIR model and a fitted SARIMA model (Chapter 10) can each tell you about the likely effect of introducing a new vaccine midway through an outbreak.*

**A:** A fitted SARIMA model can only extrapolate patterns already present in the historical case data and has no mechanism for representing a genuinely new intervention (like introducing a vaccine) that changes the underlying transmission process; a fitted SEIR model, by contrast, can explicitly simulate the effect of moving a specified fraction of the susceptible population

into a vaccinated/recovered-equivalent compartment, directly answering the counterfactual 'what would happen if we vaccinated X% of the population starting on day Y' question that the purely statistical model cannot address.

***Q4.*** *Using the worked example in Section 13.6, calculate the approximate percentage reduction in R0 achieved by the intervention scenario, and comment on whether this alone would be enough to push R0 below 1.*

**A:** R0 reduction: baseline R0 = 0.30/0.10 = 3.0; intervention R0 = 0.18/0.10 = 1.8. Percentage reduction = (3.0 " 1.8)/3.0 = 40.0%. Although this is a substantial reduction, R0 = 1.8 remains above 1, meaning the outbreak would still be expected to grow (just more slowly and to a smaller peak) under this intervention alone — additional or stronger measures would be needed to push R0 below 1 and achieve full control.

### *PYTHON — supporting code*

```
beta_baseline, beta_intervention, gamma = 0.30, 0.18, 0.10
r0_baseline = beta_baseline / gamma
r0_intervention = beta_intervention / gamma
pct_reduction = (r0_baseline - r0_intervention) / r0_baseline * 100
print(f"R0: {r0_baseline:.1f} -> {r0_intervention:.1f} ({pct_reduction:.0f}%
reduction)")
# R0: 3.0 -> 1.8 (40% reduction) -- still above 1, so the outbreak would still grow
```

***Q5.*** *In Section 13.7, explain what the fitted parameter t0 = 6.2 tells a health programme manager about the timing of the outbreak's fastest growth phase.*

**A:** t0 = 6.2 tells the programme manager that the outbreak's fastest rate of new-case accumulation is expected around week 6, which is the point at which response resources (isolation capacity, contact tracing staff, treatment supplies) are likely to be under the greatest simultaneous strain, and it also signals that growth is expected to decelerate meaningfully after this point as the outbreak approaches its eventual total size (K).

***Q6.*** *Using the Python code in Section 13.6, describe (in words) how you would modify the model to represent a vaccination campaign that moves a fraction of the susceptible population directly into the recovered compartment at day 30.*

**A:** To model a vaccination campaign moving a fraction of susceptibles to recovered at day 30, one would add a conditional step in the simulation loop: at t = 30, multiply the current S value by (1 " vaccination_coverage) and add the removed portion to R, then continue integrating the SEIR equations forward from this new starting state for the remaining time period.

***Q7.*** *A colleague argues that because the SEIR model's parameters are 'just assumptions,' its output cannot be trusted for policy decisions. Write a short response explaining the value of scenario comparison even when parameters carry uncertainty.*

**A:** A reasonable response notes that scenario comparison remains valuable under parameter uncertainty because it reveals the qualitative direction and rough magnitude of an intervention's likely impact (e.g., 'a 40% transmission reduction roughly triples the delay to peak and reduces peak size by two-thirds') even if the exact numbers shift somewhat under different plausible parameter assumptions — and running the comparison across a range of plausible parameter values (a sensitivity analysis) further strengthens confidence in the qualitative conclusion, which is often sufficient to inform a policy decision even without perfect parameter certainty.

***Q8.*** *Using Section 13.8, explain why a stochastic model might show an outbreak fizzling out even when R0 is estimated to be slightly above 1.*

**A:** In a stochastic model, each transmission event is subject to genuine chance, meaning that even when the average behaviour implied by R0 slightly above 1 would suggest growth, a run of

bad luck (an infectious individual randomly not contacting any susceptible individuals before recovering) can cause the small number of early infectious individuals to all recover without transmitting further, leading to stochastic extinction of the outbreak purely by chance — a possibility invisible to a deterministic model, which always produces the same average-behaviour trajectory regardless of chance.

***Q9.*** *In what stage of an outbreak (early, middle, or late) is the distinction between stochastic and deterministic modelling most practically important, and why?*

**A:** This distinction matters most in the EARLY stage of an outbreak, when case counts are small (a handful of individuals), because chance effects are proportionally much larger relative to a small population of infectious individuals; once an outbreak has grown to hundreds or thousands of cases, the law of large numbers means the deterministic model's average trajectory becomes an increasingly good approximation of the actual outbreak's behaviour.

***Q10.*** *Describe, in outline, how you would modify the SEIR code in Section 13.4 to introduce random variation into the number of new infections at each time step (a basic step toward a stochastic model).*

**A:** A basic stochastic version would replace the deterministic new-infection term (beta × S × I / N, applied exactly) with a randomly drawn number from an appropriate distribution (e.g., a binomial or Poisson distribution) with that same deterministic quantity as its expected value, at each discrete time step, so that the actual number of new infections varies randomly around the deterministic model's prediction rather than following it exactly.

***PYTHON — supporting code***

```
import numpy as np
def stochastic_new_infections(beta, S, I, N, rng):
    expected = beta * S * I / N
    return rng.poisson(expected)   # Poisson draw instead of the deterministic
expected value

rng = np.random.default_rng(1)
print([stochastic_new_infections(0.3, 9990, 10, 10000, rng) for _ in range(5)])
```

***Q11.*** *In the Section 13.9 case study, explain why the team investigated the discrepancy between the projected and observed peak rather than either ignoring it or immediately doubting the whole model.*

**A:** The team investigated the discrepancy, rather than dismissing or over-reacting to it, because a moderate difference between a projection and actual outcomes is expected and normal (given real-world parameter uncertainty and the possibility of subsequent interventions), and understanding WHY a discrepancy occurred is far more valuable — for both validating and improving future models — than simply noting that a discrepancy existed.

***Q12.*** *Why is it important that the team did NOT quietly revise their original projection after the fact to make it look more accurate?*

**A:** Quietly revising the original projection after the fact would misrepresent the model's true original predictive performance, undermining the value of the retrospective validation exercise entirely, since the whole point of a prospective forecast is to honestly test what the model predicted BEFORE the outcome was known, not to retroactively make the model appear more accurate than it actually was at the time.

***Q13.*** *What specific piece of real-world information (beyond the case count data itself) explained most of the discrepancy in Section 13.9?*

**A:** The specific real-world information was the timing of the water-source remediation intervention, which was implemented one week earlier (week 5) than the original response plan and baseline projection had assumed, providing a clear, well-supported explanation for both the earlier and the lower observed peak.

***Q14.*** *How would you use this retrospective validation result to argue for earlier intervention timing in a future, similar outbreak?*

**A:** This result provides concrete, quantified evidence (an estimated roughly 20% reduction in peak case count from acting one week earlier) that can be used to argue, with real supporting data rather than only theoretical modelling, for prioritising faster water-source remediation timelines in future cholera outbreak response protocols.

***Q15.*** *Explain, in your own words, why a prospective forecast (made before the outcome is known) is a more rigorous test of a model than checking a model's fit to data it was trained on.*

**A:** A prospective forecast is a more rigorous test because the model's parameters and assumptions were fixed and committed to BEFORE the actual outcome was known, eliminating any possibility of the model or its parameters being tuned, consciously or unconsciously, to match an already-observed result — checking a model's fit to its own training data, by contrast, can produce a misleadingly good-looking fit that says little about genuine predictive validity.

***Q16.*** *Calculate the herd immunity threshold for a disease with R0 = 5, using the formula in Section 13.10.*

**A:** Herd immunity threshold = 1 " 1/R0 = 1 " 1/5 = 1 " 0.2 = 0.8, meaning approximately 80% of the population must be immune to interrupt sustained transmission for a disease with R0 = 5.

***PYTHON — supporting code***

```
r0 = 5
herd_immunity_threshold = 1 - 1/r0
print(f"Herd immunity threshold: {herd_immunity_threshold:.0%}")
# Herd immunity threshold: 80%
```

***Q17.*** *Explain why the herd immunity threshold formula assumes homogeneous mixing, and why real populations often violate this assumption.*

**A:** The formula assumes homogeneous mixing because its derivation relies on every immune individual reducing transmission risk equally and uniformly across the whole population; real populations violate this because people preferentially contact household members, classmates, and geographic neighbours far more than random members of the wider population, meaning local immunity gaps can persist and sustain transmission even when the population-wide average is high.

***Q18.*** *Why can a population above the calculated herd immunity threshold still experience a limited, self-terminating outbreak following an imported case?*

**A:** A population above the herd immunity threshold can still experience a limited, self-terminating outbreak following an imported case because the threshold describes the AVERAGE condition needed to prevent SUSTAINED, ever-growing transmission across the whole population — an imported case can still infect a small number of susceptible contacts locally before the chain of transmission runs out of new susceptible individuals to infect, producing a small, finite cluster rather than either zero cases or an uncontrolled epidemic.

***Q19.*** *Explain, using Section 13.10, why a 'local pocket' of low vaccination coverage can sustain measles transmission even when a state's overall average coverage is high.*

**A:** A local pocket of low vaccination coverage can sustain transmission because, due to non-homogeneous mixing (clustering of unvaccinated individuals within specific communities, schools, or social networks), the EFFECTIVE local reproduction number within that pocket can remain above 1 even while the state's population-wide average coverage exceeds the calculated herd immunity threshold, since that average conceals substantial underlying local variation.

***Q20.*** *Why is R0 = 15 for measles associated with such a high (93%) herd immunity threshold, compared to a disease with a lower R0?*

**A:** A higher R0 means each infectious individual is expected to generate more secondary infections in a fully susceptible population, so proportionally more of the population must already be immune to bring the AVERAGE number of secondary infections down below the critical value of 1 — the herd immunity threshold formula (1 " 1/R0) mathematically increases as R0 increases, which is why measles, with its very high R0, requires such an unusually high immunity threshold compared to less transmissible diseases.

***Q21.*** *Using Section 13.11, explain one thing an agent-based model can represent that a compartmental (SEIR) model cannot.*

**A:** An agent-based model can represent realistic, heterogeneous individual contact networks (household, workplace, school-specific contact patterns) and individually targeted interventions, both of which a compartmental model's aggregate, well-mixed population assumption cannot capture at all.

***Q22.*** *Why are agent-based models not the default choice for most routine forecasting tasks, according to Section 13.11?*

**A:** Agent-based models are not the default choice for routine forecasting because they are considerably more computationally intensive to build and run, and require much

more detailed population structure and contact-pattern data to parameterise credibly, making them a poor fit for routine, resource-constrained forecasting tasks where a compartmental model's simpler structure is generally sufficient and more practical.

***Q23.*** *Propose a specific Nigerian public health scenario where an agent-based model's ability to represent individual contact networks would be particularly valuable.*

**A:** A specific scenario: modelling the spread of a highly contagious respiratory illness through a specific, densely networked urban settlement with known household and market contact patterns, where the effectiveness of a targeted, household-level quarantine strategy depends critically on the actual structure of who contacts whom — detail a compartmental model's aggregate, homogeneous-mixing assumption cannot represent.

***Q24.*** *What additional data would an agent-based model typically require that a compartmental model does not?*

**A:** An agent-based model typically requires detailed data on population structure (household composition, workplace/school assignment) and realistic contact network patterns between individuals, considerably more granular information than the aggregate case counts and simple transmission/recovery rate parameters a compartmental model requires.

***Q25.*** *Explain why 'ring vaccination around a specific detected case' is an example of an intervention better suited to agent-based than compartmental modelling.*

**A:** Ring vaccination targets specific individuals in direct contact with a detected case, which requires representing each individual's specific contact network to determine who should be vaccinated — a compartmental model's aggregate S/E/I/R counts have no concept of 'this specific person's contacts,'

making it structurally unable to represent this kind of individually-targeted intervention at all.

***Q26.*** *Summarise, in your own words, why this chapter treats mechanistic and statistical models as complementary rather than competing approaches.*

**A:** Mechanistic and statistical models are treated as complementary because they answer different kinds of questions — mechanistic models can project the consequences of a scenario that has never been observed (a proposed intervention), while statistical models extrapolate patterns already present in observed data; neither fully substitutes for the other, and a mature applied analysis often uses both together.

***Q27.*** *Explain the common thread linking the herd immunity threshold discussion (Section 13.10) and the agent-based models discussion (Section 13.11).*

**A:** The common thread is that both herd immunity thresholds and agent-based models push beyond the simplifications of the basic homogeneous-mixing SEIR framework — the herd immunity threshold concept assumes a level of population homogeneity that rarely holds exactly in practice, and agent-based models are one way of relaxing that assumption to capture heterogeneous, individual-level contact patterns explicitly.

***Q28.*** *Why does this chapter emphasise retrospectively validating a model against what actually happened (Section 13.9), rather than treating a model's initial projections as the final word?*

**A:** Retrospective validation is emphasised because a model's initial projections, made under necessarily uncertain assumptions, should be revisited against what actually happened once real outcome data becomes available — this discipline both tests the model's credibility and provides valuable lessons for improving future model specifications.

***Q29.*** *Propose a scenario from your own field of interest where an agent-based model might capture something a compartmental SEIR model would miss.*

**A:** Answers will vary, but a strong example would involve substantial heterogeneity in contact patterns (e.g., school-based transmission, or transmission concentrated in specific social networks) that a homogeneous-mixing compartmental model would poorly represent, but that an agent-based model with explicit individual contact structures could capture.

***Q30.*** *If you had to explain the practical limits of the herd immunity threshold concept to a policymaker, what would you say?*

**A:** A reasonable explanation: herd immunity thresholds are calculated under an assumption of homogeneous mixing and constant, uniform immunity, but in reality, immunity and contact patterns vary substantially across subgroups and can wane over time, meaning the true practical threshold for sustained control may differ meaningfully from the simple textbook calculation.

***Q31.*** *Using Section 13.13, explain why the homogeneous mixing assumption in a standard SEIR model might be a poor fit for HIV transmission.*

**A:** HIV transmission occurs specifically through sexual and needle-sharing contact networks rather than through random, generalised population mixing, meaning an individual's actual risk depends heavily on their specific network position and behaviour, a structure a homogeneous-mixing SEIR model, which assumes every individual is equally likely to contact every other individual, cannot represent.

***Q32.*** *Describe what a 'super-spreader' individual represents in a network transmission model, and why a compartmental model cannot represent this concept directly.*

**A:** A 'super-spreader' individual is someone with an unusually large number of contacts (a highly connected node in the transmission network) who can, through their own single infection, generate a disproportionately large number of secondary infections; a compartmental model, which only tracks aggregate counts within each state and has no representation of individual-level contact structure, has no mechanism to represent such heterogeneity in individual connectivity.

***Q33.*** *Propose a Nigerian disease context where network structure (household, occupational, or social) would likely matter a great deal for transmission dynamics.*

**A:** A reasonable example: tuberculosis transmission within densely connected household or workplace networks, where sustained close contact substantially elevates transmission risk compared to brief, casual community contact that a homogeneous-mixing model would treat as equally risky.

***Q34.*** *What additional data would be needed to build a network transmission model that would not be needed for a standard SEIR model?*

**A:** A network model would require detailed contact network data — who is connected to whom, and with what frequency or intensity of contact — which is considerably more detailed, harder to collect, and more resource-intensive to gather than the aggregate population-level parameters (transmission rate, recovery rate) a standard SEIR model requires.

***Q35.*** *How might the protective effect of network structure described in Section 13.13 relate to the herd immunity threshold discussed earlier (Section 13.10)?*

**A:** Network structure can create a natural analogue to herd immunity by limiting how far an early cluster of cases can spread through the network before running out of new susceptible contacts within its local network neighbourhood, potentially

allowing an outbreak to remain contained even at an overall population immunity level below the classical herd immunity threshold calculated under a homogeneous-mixing assumption.

## Chapter 14

## Spatial Epidemiology and Disease Mapping

***Q1.*** *Explain why a choropleth map of raw leishmaniasis case counts by state could be misleading, and describe how you would correct it.*

**A:** A choropleth map of raw case counts largely reproduces the underlying population distribution (more populous states will tend to have more cases simply because they have more people, regardless of true risk), which could misleadingly suggest that the most populous states are the 'worst affected' when their actual risk per capita may be lower than a smaller state's; the correction is to map a rate (cases per population, e.g., per 100,000) rather than the raw count.

***Q2.*** *What does a positive, statistically significant Moran's I for malaria TPR across LGAs suggest about the underlying drivers of transmission?*

**A:** A positive, statistically significant Moran's I suggests that high-TPR LGAs are non-randomly located near other high-TPR LGAs (spatial clustering), which is consistent with shared environmental or ecological drivers of malaria transmission (such as rainfall, vegetation, or vector habitat) operating at a spatial scale larger than a single LGA, rather than each LGA's TPR being an independent, locally-determined outcome.

***Q3.*** *Why might a BYM (Bayesian small-area) model produce a more reliable risk estimate for a low-population LGA than the raw observed rate?*

**A:** A BYM model 'borrows strength' from the surrounding, spatially connected LGAs when estimating a low-population LGA's risk, shrinking its potentially very noisy raw rate toward a more stable regional estimate; since a raw rate calculated

from a small number of cases and a small population can swing enormously based on just one or two additional/fewer cases, this smoothing produces a more reliable and less erratic estimate for small areas.

***Q4.*** *Using the worked example in Section 14.5, explain in your own words what the permutation test p-value of 0.001 is actually testing (i.e., what null hypothesis is being compared against).*

**A:** The permutation test is testing the null hypothesis of complete spatial randomness — that the observed TPR values, if randomly reshuffled across the same fixed spatial arrangement of LGAs many times (999 times here), would produce a Moran's I as extreme as 0.42 purely by chance; a p-value of 0.001 means only about 1 in 1,000 such random reshufflings produced an I that extreme, providing strong evidence against pure spatial randomness.

***PYTHON — supporting code***

```
observed_moran_i = 0.42
n_permutations = 999
n_as_extreme = 0   # none of the 999 random permutations reached I=0.42
p_value = (n_as_extreme + 1) / (n_permutations + 1)
print(f"Permutation p-value: {p_value:.3f}")
```

***Q5.*** *In Section 14.6, explain why LGA C's raw and smoothed rates are so close together, while LGA A and B's rates change substantially under smoothing.*

**A:** LGA C's raw and smoothed rates are close together because its much larger population (150,000) gives its raw rate greater inherent statistical stability/reliability, so the empirical Bayes method places more weight on (has less need to correct) its own observed data; LGAs A and B, with far smaller populations and correspondingly noisier raw rates, are given much less weight relative to the regional average, so their smoothed estimates are pulled substantially toward it.

***PYTHON — supporting code***

```
cases_a, pop_a = 2, 8000
cases_c, pop_c = 18, 150000
raw_rate_a = cases_a / pop_a * 100000
raw_rate_c = cases_c / pop_c * 100000
print(f"LGA A raw rate: {raw_rate_a:.1f} (small pop -> unstable)")
print(f"LGA C raw rate: {raw_rate_c:.1f} (large pop -> stable, barely shrinks
under smoothing)")
```

***Q6.*** *Using the table in Section 14.6, if LGA A's population were instead 80,000 (ten times larger) with the same 2 cases, would you expect its smoothed rate to move further from or closer to its raw rate compared to the original scenario? Explain why.*

**A:** If LGA A's population were 80,000 (ten times larger) with the same 2 cases, its raw rate would already be much lower (2.5 per 100,000) and, more importantly, its increased population size would give the raw rate greater inherent reliability, so the smoothed estimate would move much less from the raw rate than in the original small-population scenario — the smoothing 'trusts' larger-population estimates more.

***PYTHON — supporting code***

```
# Empirical Bayes smoothing sensitivity to population size (Section 14.6)
def smooth_rate(cases, population, regional_avg, weight_regional):
    raw_rate = cases / population * 100000
    return weight_regional * regional_avg + (1 - weight_regional) * raw_rate

regional_avg = 13.5
small_pop = smooth_rate(cases=2, population=8000, regional_avg=regional_avg,
weight_regional=0.6)
large_pop = smooth_rate(cases=2, population=80000, regional_avg=regional_avg,
weight_regional=0.15)
print(f"Small population (8,000) smoothed: {small_pop:.1f}")
print(f"Larger population (80,000) smoothed: {large_pop:.1f}  -- moves much less
from its raw rate")
```

***Q7.*** *A journalist wants to publish a map showing 'the most dangerous LGA for leishmaniasis' based on raw case rates alone. Write a short response explaining the risk of doing this without smoothing, using a concrete hypothetical example.*

**A:** A response should explain that raw rates from small LGAs are statistically unreliable and can be extreme purely by chance (as illustrated by LGA A and B in Section 14.6), so declaring 'the most dangerous LGA' based on unsmoothed rates risks incorrectly stigmatising a small-population area that simply had a couple of unlucky cases, while potentially overlooking a genuinely higher-risk but larger, more stably-estimated area; smoothing the rates first would give a more defensible basis for such a claim.

***Q8.*** *Using Section 14.7, identify two ways a facility name mismatch between a case dataset and a facility registry could cause a spatial analysis to silently produce incorrect results.*

**A:** A facility name mismatch could cause cases to be silently dropped from a spatial analysis entirely (if the join fails to match them to any facility record) or, worse, silently matched to the WRONG facility with a similar name, attributing cases to an incorrect location without any visible error message — both would corrupt the resulting spatial pattern without necessarily producing any obvious warning sign.

***Q9.*** *Why is a wrong hemisphere sign convention (e.g., a positive instead of negative latitude) a particularly dangerous kind of geocoding error to miss?*

**A:** A wrong hemisphere sign convention is particularly dangerous because it does not produce an obviously invalid coordinate (the resulting point is still a technically valid latitude/longitude on Earth), so it can pass basic range-validity checks while placing the point in a completely wrong location — for Nigeria, this could place a facility incorrectly in the Atlantic Ocean or on a different continent, a serious but easily-missed error if the resulting map is not visually inspected.

***Q10.*** *How would you check whether an administrative boundary shapefile you are using is out of date relative to recent LGA changes?*

**A:** One would check the file's stated publication or last-update date against known recent LGA boundary changes or creations, compare the total number of administrative units in the shapefile against the current officially recognised count, and, ideally, cross-reference a sample of unit names and boundaries against an authoritative current source (such as the National Population Commission or NBS) before relying on the file for analysis.

***Q11.*** *In the Section 14.8 case study, explain why the ward-level map revealed information the LGA-level map had concealed.*

**A:** The ward-level map revealed that the apparent LGA-wide stunting hotspot was actually driven by just two of eleven wards, with the remaining nine wards showing typical prevalence — information that was entirely averaged away and invisible in the coarser LGA-level aggregation.

***Q12.*** *Why would an intervention based only on the LGA-level map have been less efficient than one based on the ward-level map?*

**A:** An LGA-wide intervention would spread supplementary nutrition resources across all eleven wards, including nine that do not actually need them, diluting the resources available for the two wards with genuinely elevated prevalence and therefore delivering a less effective, less efficient intervention than one specifically targeted at the two high-burden wards.

***Q13.*** *At what point would going to an even finer geographic resolution (e.g., individual settlements) stop being useful, and why?*

**A:** At some point, going to an even finer resolution (such as individual settlements) would produce estimates based on such

small sample sizes that they become highly statistically unreliable (echoing Chapter 14's small-area smoothing discussion), so the usefulness of finer resolution is limited by the point at which the underlying survey or case data can no longer support a stable, reliable estimate at that resolution.

***Q14.*** *Propose one practical constraint that might prevent a health programme from always using the finest available geographic resolution for intervention targeting.*

**A:** A practical constraint is that finer-resolution data may simply not be available or may be prohibitively expensive/time-consuming to collect (survey samples are often only designed to be representative down to a certain administrative level), meaning a programme may be limited to the finest resolution its available data can actually support reliably.

***Q15.*** *How would you decide, in general, which geographic resolution is 'fine enough' for a specific health indicator and intervention decision?*

**A:** The appropriate resolution is generally the finest level at which the available data still produces reasonably stable, reliable estimates (checked, for example, via the confidence interval width or minimum sample size at that level) — going finer than this trades genuine precision for noise, while staying coarser than necessary risks concealing actionable heterogeneity, as illustrated in Section 14.8.

***Q16.*** *Using Section 14.9, explain why an LGA directly on Nigeria's northern border might show an apparently unexplained spatial pattern in a Lassa fever analysis.*

**A:** An LGA directly on Nigeria's northern border might show an apparently isolated, unexplained pattern in a Lassa fever analysis because the analysis has no data on rodent populations, human movement, or case incidence occurring just across the border, potentially truncating a single, continuous cross-border

transmission zone into what appears, within the Nigeria-only dataset, as an isolated and unexplained cluster.

***Q17.*** *Propose one practical step a state or national programme could take to reduce the edge effect problem described in Section 14.9.*

**A:** A practical step: establishing a formal, even if limited, data-sharing arrangement or informal information exchange with health authorities in the neighbouring country's border region, allowing at least aggregate case count and timing information to be incorporated into a more complete spatial picture for these specific border-adjacent analyses.

***Q18.*** *Why might treating a state boundary as an epidemiological barrier be especially problematic for a disease like cholera, compared to a disease with very limited geographic mobility?*

**A:** For cholera, which can spread through water sources and human movement that plausibly cross porous borders and river systems without any regard for administrative boundaries, treating a state boundary as an epidemiological barrier is especially misleading, since the disease's actual transmission dynamics are likely continuous across that boundary in a way that a boundary-and-mobility-limited disease would not be.

***Q19.*** *How would you flag a border-adjacent LGA's spatial analysis result as having 'incomplete spatial context,' in a technical report?*

**A:** You would add an explicit caveat in the technical report's methods or limitations section, such as: 'Analysis for LGA X, located on Nigeria's border with [country], may be affected by unobserved cross-border transmission dynamics not captured in this dataset; results for this and similarly situated border LGAs should be interpreted with this limitation in mind.'

***Q20.*** *Explain why cross-border data-sharing arrangements, while logistically difficult, would directly address the root cause of the edge effect problem.*

**A:** Cross-border data-sharing arrangements would directly address the root cause because the edge effect problem stems specifically from a genuine gap in data VISIBILITY across the border, not from any flaw in the analytic method itself — providing the missing cross-border data (even in aggregate form) would allow the same spatial methods (Chapter 14) to be applied to a more complete, continuous picture of the true transmission zone.

***Q21.*** *Using Section 14.10, propose one specific way a vegetation index could serve as a proxy for vector breeding habitat suitability in a leishmaniasis spatial model.*

**A:** A vegetation index could proxy for the presence of standing water, shaded vegetation, or specific soil moisture conditions favourable to the sandfly vector responsible for leishmaniasis transmission, allowing a model to test whether areas with vegetation patterns consistent with suitable vector habitat show correspondingly higher case rates, independent of population density alone.

***Q22.*** *Why might nighttime lights data be a useful substitute for outdated census population figures in a rapidly urbanising area?*

**A:** Nighttime lights data can serve as a real-time, frequently-updated proxy for population density and urbanisation extent in areas where the official census figures are outdated or where rapid, unplanned urban growth has occurred since the last census, providing a more current population signal than an ageing census-based projection alone.

***Q23.*** *What additional technical skill does incorporating remote sensing covariates require, beyond the vector/tabular spatial methods emphasised elsewhere in Chapter 14?*

**A:** Incorporating remote sensing covariates typically requires skill in processing raster (grid-cell-based) geospatial data

formats, distinct from the vector (point/polygon-based) geospatial data handling emphasised in this chapter's core choropleth mapping and Moran's I examples, often using specialised raster-processing libraries.

***Q24.*** *Explain one potential benefit of incorporating environmental covariates for an under-surveyed area with limited case-count data.*

**A:** For an under-surveyed area with limited case-count data, environmental covariates derived from satellite data (which are available globally and consistently, regardless of local surveillance capacity) can help a spatial model make a more informed prediction of likely disease risk even where direct case-count evidence is sparse or unreliable.

***Q25.*** *Propose a specific research question that could be tested by combining case-count data with a satellite-derived rainfall covariate.*

**A:** A specific research question: does higher satellite-derived rainfall in the preceding one to two months predict higher subsequent malaria case rates at the LGA level, after controlling for population density — directly testing the well-established rainfall-transmission relationship using an independent, environmentally-grounded covariate rather than relying on calendar month alone as a proxy for rainfall season.

***Q26.*** *Summarise, in your own words, why this chapter treats the modifiable areal unit problem as a fundamental, not merely technical, limitation of spatial analysis.*

**A:** The modifiable areal unit problem is treated as fundamental because it means the very pattern an analyst observes and reports can change substantially depending on an essentially arbitrary choice of administrative boundaries for aggregation — this is not a fixable technical error but an inherent limitation of any areally-aggregated spatial analysis that must be explicitly acknowledged.

***Q27.*** *Explain the common thread linking the edge-effects discussion (Section 14.9) and the moving-hotspot case study (Section 14.8).*

**A:** The common thread is that both edge effects and the moving-hotspot case study illustrate how a spatial pattern's apparent shape and location can be an artefact of how the analysis boundary or aggregation unit was drawn, rather than reflecting the true, boundary-independent underlying spatial process of disease risk.

***Q28.*** *Why does this chapter recommend incorporating remote sensing and environmental covariates (Section 14.10) rather than relying on case counts alone?*

**A:** Remote sensing and environmental covariates are recommended because they provide a continuous, boundary-independent measure of the environmental drivers of disease risk (vegetation, rainfall, temperature), which can explain and help interpret a spatial pattern in a way that raw administrative-unit case counts, which are inherently tied to an arbitrary boundary choice, cannot.

***Q29.*** *Propose a spatial analysis from your own field where edge effects at an administrative or national boundary (Section 14.9) could distort a naive interpretation.*

**A:** Answers will vary, but should identify a specific cross-border or cross-LGA spatial analysis where cases or environmental exposure near the boundary could be miscounted or mis-attributed due to the artificial administrative cutoff, distorting the apparent local risk pattern.

***Q30.*** *If you had to explain to a journalist why 'the most dangerous LGA' can change depending on how the map is drawn, what would you say?*

**A:** A reasonable explanation: because a choropleth map's visual pattern depends partly on the arbitrary choice of administrative

units used to aggregate the data, redrawing the same underlying case data at a different level of aggregation (LGA versus ward, for example) can genuinely change which area appears 'most affected,' a limitation inherent to any areally-aggregated map.

***Q31.*** *Using Section 14.12, explain why a spatiotemporal model can produce more stable small-area estimates than either a purely spatial or purely temporal model alone.*

**A:** A spatiotemporal model produces more stable estimates because it can borrow statistical strength from both nearby locations in the same time period AND from the same location across nearby time periods simultaneously, giving a small-area, single-month estimate more supporting information than either a purely spatial (single time period, many locations) or purely temporal (single location, many time periods) approach could provide alone.

***Q32.*** *Describe a Nigerian surveillance dataset where a spatial hotspot might plausibly shift location from one season to the next.*

**A:** A plausible example: a vector-borne disease hotspot that shifts from one LGA to a neighbouring LGA as seasonal rainfall patterns and vector breeding sites change from the early to the late rainy season within the same year.

***Q33.*** *Why does fitting a spatiotemporal model typically require more specialised software than fitting a purely spatial model?*

**A:** A spatiotemporal model requires jointly modelling correlation structures across both space and time simultaneously, which involves a substantially larger and more complex set of parameters and computational demands than either a purely spatial or purely temporal model, generally requiring specialised software such as R's INLA package to fit within a reasonable computation time.

***Q34.*** *What is meant by 'borrowing strength' in both the spatial and temporal dimensions simultaneously?*

**A:** Borrowing strength in both dimensions means a specific LGA-month's smoothed estimate is informed by data from its geographic neighbours in the same month AND by its own data from adjacent months, combining both sources of related information rather than relying on only one.

***Q35.*** *How does the added complexity of a spatiotemporal model relate to the sensitivity-analysis discipline discussed in Section 13.12?*

**A:** Both reflect the same underlying discipline of being explicit about a model's assumptions and their practical consequences — sensitivity analysis (Section 13.12) tests how much a mechanistic model's conclusions depend on uncertain parameter choices, while the added complexity of a spatiotemporal model should likewise be evaluated against whether it meaningfully improves practical decision-relevant estimates over simpler alternatives.

## Chapter 15

## Big Data, AI, and Deep Learning in Health Systems

***Q1.*** *Give an example of each of the 'three Vs' (volume, velocity, variety) from a Nigerian health data context you are familiar with.*

**A:** Examples (any three Vs illustrated with a Nigerian example are acceptable): Volume — a nationwide HMIS transaction-level database too large to load into a single spreadsheet; Velocity — a real-time IDSR/SORMAS surveillance dashboard requiring continuous streaming updates rather than a monthly batch upload; Variety — a hospital's records combining structured lab results, free-text clinician notes, and radiology images that must be integrated for a single patient's record.

***PYTHON — supporting code***

```
volume_examples = {"HMIS transaction logs": "millions of rows/month"}
velocity_examples = {"IDSR live dashboard": "continuous weekly updates"}
variety_examples = {"Hospital EHR": "structured labs + free-text notes +
images"}
for v in [volume_examples, velocity_examples, variety_examples]:
    print(v)
```

***Q2.*** *Explain why a CNN trained only on high-quality reference blood smear images might perform poorly when deployed using smartphone-captured images in a rural facility.*

**A:** A CNN trained only on high-quality reference images learns to recognise patterns specific to that consistent image quality, staining technique, and equipment; smartphone-captured images from a rural facility typically have more variable lighting, focus, colour balance, and resolution, which the model was never exposed to during training — so its learned filters may not correctly generalise to this different, noisier input distribution, causing accuracy to drop, sometimes substantially, in the field deployment setting.

***Q3.*** *A colleague proposes using a deep neural network for a study of 300 patients and 12 clinical variables. What would you advise, and why?*

**A:** For a study of 300 patients and 12 variables, a deep neural network is very likely to overfit given the small sample size relative to the number of parameters a network would typically have; a classical method such as logistic regression, LASSO, or a random forest (Chapters 8 and 11) would very likely perform as well or better, train much faster, and remain far more interpretable — deep learning should be reserved for cases with much larger data volumes or image/text formats, per Section 15.5.

***Q4.*** *Using the worked example in Section 15.6, explain why a CNN trained on images from a single laboratory might not generalise well to images from a different facility, even if test accuracy on the original dataset is high.*

**A:** A CNN trained on images from a single laboratory learns patterns tied to that lab's specific staining protocol, microscope optics, and lighting conditions, which may not be representative of a different facility's equipment and technique; even with high test accuracy on held-out images from the SAME laboratory, this says nothing about performance on the different image characteristics of another facility, echoing the same generalisation risk raised for grouped cross-validation in Chapter 11.

***Q5.*** *In Section 15.7, explain one advantage and one limitation of a rule-based keyword system compared to a more sophisticated NLP model, for the specific task of verbal autopsy triage.*

**A:** Advantage: a rule-based system is fully transparent and auditable — a physician reviewer can see exactly which keywords triggered a flag and judge whether the logic makes clinical sense, and it requires no large labelled training dataset to build. Limitation: it will miss cases described using synonyms, unusual phrasing, or indirect descriptions not included in its keyword list, and it cannot capture more nuanced or context-dependent linguistic patterns that a more sophisticated NLP model might detect.

***Q6.*** *Modify the Python code in Section 15.7 (in words or code) to add a third keyword category for a probable cause of death of your choosing, and describe what keywords you would include and why.*

**A:** Answers will vary depending on the chosen cause of death, but should include a plausible keyword list (e.g., for a suspected

postpartum haemorrhage category: ‘heavy bleeding after delivery’, ‘bleeding after childbirth’, ‘excessive blood loss’) and should explain the choice by referencing observable, commonly-used descriptive phrases a lay respondent might use in a verbal autopsy interview.

***PYTHON — supporting code***

```
hepatic_keywords = ["yellow eyes", "jaundice", "dark urine"]
obstetric_keywords = ["heavy bleeding after delivery", "bleeding after childbirth"]

def flag_probable_cause(narrative):
    text = narrative.lower()
    flags = []
    if any(k in text for k in obstetric_keywords):
        flags.append("possible_postpartum_haemorrhage")
    if any(k in text for k in hepatic_keywords):
        flags.append("possible_hepatic")
    return flags or ["unclassified_needs_review"]

print(flag_probable_cause("She had heavy bleeding after delivery and became
very weak."))
```

***Q7.*** *A donor wants to fund ‘an AI system’ for a Nigerian disease surveillance programme without specifying the task. Using this chapter’s arguments, write a short set of clarifying questions you would ask before recommending any specific tool.*

**A:** Reasonable clarifying questions include: What specific task should the AI system perform (image classification, text extraction, tabular risk prediction, outbreak detection)? What volume and format of data is actually available to train and validate it? Who will use its output, and what level of interpretability do they need to trust and act on it? Answering these first (per Section 15.5’s reasoning) determines whether deep learning, classical machine learning, or a simple rule-based tool is actually the appropriate recommendation.

***Q8.*** *Using Section 15.8, explain why transfer learning could make deep learning more feasible for a Nigerian research group with only 800 labelled clinical images.*

**A:** Transfer learning could make deep learning feasible with only 800 labelled images because the network does not need to learn general visual features (edges, textures, basic shapes) from scratch using those 800 images alone — it starts from a network that has already learned these general features from a much larger, unrelated image dataset, and only needs the smaller local dataset to learn the specific final distinction (e.g., parasitised vs. uninfected) built on top of those already-learned general features.

***Q9.*** *What kind of general-purpose visual features does a pre-trained network already know, and why does re-using them save the need for a much larger local training dataset?*

**A:** A pre-trained network already knows general-purpose low- and mid-level visual features such as edge detectors, colour/texture patterns, and basic shape detectors, which are broadly useful across many different image classification tasks; re-using them means the smaller local dataset only has to teach the network the final, task-specific distinction rather than the much larger and more data-hungry task of learning all visual features from zero.

***Q10.*** *Would transfer learning fully solve the generalisation concern raised in Section 15.6 about a model trained on a single laboratory's images? Explain why or why not.*

**A:** Transfer learning would not fully solve the generalisation concern, because the model's final, task-specific layers are still trained (and therefore specialised) on the particular local dataset's staining technique, microscope, and imaging conditions; the model's performance could still degrade on images from a different facility with different equipment or

technique, even though the underlying general visual features it started from are more robust than if trained from scratch on the small local dataset alone.

***Q11.*** *In the Section 15.9 case study, explain why the tool's drop in performance was specifically linked to code-switched narratives rather than a general decline in accuracy.*

**A:** The tool's performance dropped specifically for code-switched narratives because its keyword list was built entirely from formal English medical and lay terminology, and code-switched narratives frequently use local-language terms for the same symptoms that simply do not appear anywhere in that English-only list, causing the rule-based logic to systematically miss cases it would have correctly flagged had the English term been used instead.

***Q12.*** *Why might this kind of error disproportionately affect exactly the communities whose verbal autopsy data is most valuable?*

**A:** This error disproportionately affects communities where code-switching (or reliance on local-language terms) is most common in everyday speech, which likely correlates with rural or lower-literacy areas — precisely the areas where facility-based cause-of-death data is often least available, meaning verbal autopsy data (and therefore accurate tool performance) matters most in exactly the populations the tool serves least well.

***Q13.*** *Describe the process you would use to identify local-language terms to add to the tool's keyword list, drawing on Section 15.9's proposed remedy.*

**A:** You would work directly with the field interview teams who actually conduct these interviews in the affected communities, asking them to identify the specific local-language terms and phrases commonly used for each key symptom category, and

incorporate these terms directly into the tool's keyword list alongside the original English-language terms.

***Q14.*** *Would you expect a transformer-based NLP model (Section 15.4) to be more robust to this specific problem than the rule-based tool? Explain your reasoning, including any new risks it might introduce.*

**A:** A transformer-based NLP model, if trained or fine-tuned on a sufficiently large and representative sample of code-switched narratives, could plausibly handle this variation more robustly than a fixed rule-based keyword list; however, it would introduce new risks of its own — reduced transparency/ auditability for physician reviewers (Section 15.4-15.5's interpretability concerns) and a need for a much larger labelled training dataset, which may not be readily available for this specific, locally-varied linguistic context.

***Q15.*** *Why is ongoing deployment monitoring, not just initial validation, necessary to catch a problem like the one in Section 15.9?*

**A:** Ongoing monitoring is necessary because a tool's real-world performance can degrade in ways that only become visible once it is deployed across the full diversity of the actual target population — a validation conducted on a linguistically non-representative initial sample (as in Section 15.9) cannot by itself reveal a problem that only manifests in a different subgroup encountered later during broader deployment.

***Q16.*** *Using Section 15.10, explain why a lower classification threshold (favouring sensitivity) might be appropriate for a blood smear screening tool meant to flag samples for expert review.*

**A:** A lower classification threshold favouring sensitivity is appropriate for a blood smear screening tool because the cost of a false negative (a missed, truly parasitised sample) is a

genuine, potentially serious clinical miss, while the cost of a false positive is only a few minutes of an expert reviewer's time re-examining an uninfected sample — an asymmetric cost structure that justifies deliberately shifting the threshold to catch more true positives at the expense of more false alarms.

***Q17.*** *Propose a clinical scenario where the opposite trade-off (favouring specificity over sensitivity) would instead be more appropriate, and explain why.*

**A:** A scenario favouring specificity over sensitivity: a screening tool used to decide whether to initiate a costly, invasive, or higher-risk confirmatory procedure directly (rather than merely flagging for a quick expert review), where a false positive would trigger significant unnecessary cost, risk, or patient anxiety, making it more important to minimise false alarms even at some cost to catching every true case immediately.

***Q18.*** *Why should the choice of classification threshold be made 'in consultation with the clinical team,' rather than left to a data scientist working alone?*

**A:** This choice should involve the clinical team because they, not the data scientist alone, understand the real-world operational and clinical consequences of a missed case versus an unnecessary review or procedure in their specific setting, and the appropriate threshold is fundamentally a judgment about acceptable trade-offs in patient care and resource use, not a purely statistical or technical decision.

***Q19.*** *Explain, in your own words, why a default threshold of 0.5 has no inherent clinical meaning.*

**A:** A default threshold of 0.5 simply reflects the midpoint of the model's raw probability output scale and has no inherent connection to the actual relative clinical costs of a false positive versus a false negative in any specific application — using it by default, without deliberate consideration, is essentially an

arbitrary and unexamined choice rather than a clinically justified one.

***Q20.*** *If false negatives and false positives had genuinely equal costs in a given application, would adjusting the threshold away from 0.5 still be justified? Explain your reasoning.*

**A:** Even with genuinely equal costs, adjusting the threshold away from 0.5 could still be justified if the two classes (positive and negative cases) occur at very different rates (class imbalance) in the deployment population, since a threshold optimised for balanced classes may not be optimal under substantial class imbalance — though in the specific case of truly equal costs and balanced classes, 0.5 would indeed often be a reasonable, well-justified choice.

***Q21.*** *Using Section 15.11, explain why training a large neural network from scratch could be a poor institutional fit for many Nigerian research settings, beyond the risk of overfitting discussed in Section 15.5.*

**A:** Beyond the risk of overfitting on a small dataset, training a large neural network from scratch requires substantial, often specialised computing hardware (GPUs) and reliable electricity access that may be genuinely difficult or expensive to secure consistently in many Nigerian institutional settings, making it a poor practical fit regardless of the dataset's size or the model's theoretical suitability.

***Q22.*** *Why does even RUNNING (not training) a large pre-trained model impose an ongoing resource burden, according to Section 15.11?*

**A:** Even running (not training) a large pre-trained model for routine inference still requires meaningful computational resources (memory, processing power, and associated electricity costs) every time the model is used, imposing an ongoing, recurring resource burden on an institution rather than a one-time training cost alone.

***Q23.*** *Propose one way a resource-constrained institution could still benefit from deep learning methods without needing to train a large model from scratch.*

**A:** A resource-constrained institution could use transfer learning (Section 15.8), starting from a smaller, more efficient pre-trained model and fine-tuning only its final layers on local data, which requires substantially less computing power than training a large model entirely from scratch, or could use cloud-based inference services that shift the computational burden elsewhere, if budget and data-sharing considerations permit.

***Q24.*** *Explain the connection between Section 15.11's resource-cost argument and Section 15.8's discussion of transfer learning.*

**A:** Section 15.11's resource-cost argument reinforces Section 15.8's transfer learning discussion by providing an additional, practical (not just statistical/data-volume) reason to prefer transfer learning over training from scratch — transfer learning is not only more data-efficient but also considerably less computationally demanding, addressing both of the major practical barriers to deep learning adoption in resource-constrained settings.

***Q25.*** *Why is 'a tool intended for sustained, in-country institutional use' an important framing when choosing between a classical model and a large deep learning model?*

**A:** This framing matters because a tool intended for ONE-OFF research demonstration purposes can reasonably absorb a higher, temporary resource cost, whereas a tool intended for ONGOING institutional use needs to be sustainable within the institution's regular, long-term computing budget and infrastructure — a distinction that should directly inform the choice between a resource-intensive deep learning model and a more modest classical alternative.

***Q26.*** *Summarise, in your own words, why this chapter argues that matching the tool to the task and data scale matters more than always choosing the most advanced available method.*

**A:** This chapter argues for matching the tool to the task and data scale because a more advanced method (such as deep learning) is not universally superior — it earns its additional complexity, data requirements, and reduced interpretability only in specific circumstances (large datasets, image/text data), and applying it by default to an unsuitable task wastes resources and can even underperform a simpler, well-chosen alternative.

***Q27.*** *Explain the common thread linking the code-switched text case study (Section 15.9) and the environmental cost of large AI models discussion (Section 15.11).*

**A:** The common thread is that both examples illustrate hidden, easily-overlooked limitations of AI tools that only become apparent under real-world deployment conditions — an NLP tool's blind spot for code-switched text, and a large AI model's often-unstated environmental and computational resource cost — both of which should be considered before, not after, deployment.

***Q28.*** *Why does this chapter insist on explicitly considering the relative cost of false positives versus false negatives (Section 15.10) before deploying an AI-assisted triage tool?*

**A:** Explicitly weighing the cost of false positives against false negatives matters because these costs are rarely symmetric in a clinical triage context — a missed severe case (false negative) may carry a far higher human cost than an unnecessary referral (false positive), and this asymmetry should directly inform how a model's decision threshold is set, rather than defaulting to a generic 0.5 probability cutoff.

***Q29.*** *Propose an AI application from your own field where the cost of a false negative would clearly outweigh the cost of a false positive, or vice versa, and explain the practical implication.*

**A:** Answers will vary, but a strong example involves a serious, time-sensitive condition (where a missed case, a false negative, could be life-threatening) contrasted with a low-stakes screening context (where a false positive mainly causes minor inconvenience or unnecessary follow-up testing).

***PYTHON — supporting code***

```
cost_false_negative = "missed severe case -- potentially life-threatening"
cost_false_positive = "unnecessary referral -- inconvenience, some cost"
print("Asymmetric costs should shift the decision threshold away from the default
0.5")
```

***Q30.*** *If you had to advise a Nigerian institution considering a large, resource-intensive AI model, what questions from this chapter would you want them to answer first?*

**A:** Reasonable questions include: What specific problem and data volume/format is this model meant to address? Has a simpler baseline method already been tried and found insufficient? What is the relative cost of a false positive versus a false negative in this specific application, and does the chosen model's threshold reflect that asymmetry?

***Q31.*** *Using Section 15.13, explain how federated learning addresses the data sovereignty concerns raised in Chapter 16.*

**A:** Federated learning addresses data sovereignty concerns because each institution's raw patient data never leaves that institution's own servers or jurisdiction — only aggregated model updates (not identifiable patient records) are shared centrally, directly avoiding the cross-border or cross-institutional data transfer concerns raised in Chapter 16's discussion of data sovereignty.

***Q32.*** *Describe what is shared between institutions in a federated learning setup, and, just as importantly, what is NOT shared.*

**A:** What is shared: the model's learned parameters or gradient updates from each institution's local training round. What is NOT shared: the underlying raw patient records, images, or any directly identifiable data, which remain physically and legally within each contributing institution at all times.

***Q33.*** *Why might a model trained via federated learning across several institutions generalise better than a model trained on a single institution's data alone?*

**A:** A federated model can generalise better because it effectively learns from the combined statistical patterns of multiple institutions' more diverse patient populations, equipment, and case mixes, rather than being limited to (and potentially overfit to) a single institution's more narrow and specific data characteristics, directly addressing the single-facility generalisation concerns raised in Sections 11.5 and 15.6.

***Q34.*** *What practical or technical challenges might arise in coordinating a federated learning project across multiple Nigerian teaching hospitals?*

**A:** Practical challenges include coordinating consistent data formats and variable definitions across institutions (echoing Chapter 2's interoperability concerns), ensuring reliable and secure network infrastructure for the periodic exchange of model updates, and establishing clear governance agreements about how the resulting shared model may subsequently be used and by whom.

***Q35.*** *How does federated learning relate to the subgroup generalisation concerns raised in Sections 11.5 and 11.7?*

**A:** Federated learning directly targets the same underlying problem raised in Sections 11.5 and 11.7 — that a model trained and validated on a single institution's data may not generalise well elsewhere — by allowing the model to learn

from multiple institutions' diverse data during training itself, rather than only discovering the generalisation gap after deployment via a subgroup audit.

## Chapter 16

## Data Ethics, Governance, and the Future of Health Data Science in Nigeria

***Q1.*** *A dataset removes patient names but retains exact age, LGA, and a rare diagnosis. Explain the re-identification risk this creates and propose a mitigation.*

**A:** Even without a name, the combination of exact age, LGA, and a rare diagnosis could be sufficient to identify a specific individual, particularly in a small community where such a diagnosis is uncommon and widely known locally; mitigation includes coarsening age into wider bands, aggregating geography to a larger administrative unit, and/or applying a formal k-anonymity check (Section 16.5) to ensure every combination of these quasi-identifiers is shared by at least several individuals before release.

***PYTHON — supporting code***

```
quasi_identifiers_example = {"age": "exact", "lga": "specific", "diagnosis": "rare"}
mitigation = "coarsen age into bands, aggregate to a wider geographic unit, or
suppress the rare-diagnosis flag"
print(mitigation)
```

***Q2.*** *Describe a plausible scenario in which a clinical risk-prediction model trained on Nigerian tertiary hospital data could disadvantage patients from primary health centres.*

**A:** If historical access to specialist referral or advanced diagnostics was concentrated at tertiary hospitals, a model trained on outcomes from that setting may learn patterns that do not reflect (and may be poorly calibrated for) the different case mix and available information at primary health centres,

meaning patients at PHCs — who may already have less access to specialist resources — could receive systematically less reliable risk predictions if the model is deployed there uncritically, compounding an existing access inequity.

***Q3.*** *What provisions would you want to see in a data-sharing agreement before contributing a Nigerian patient cohort dataset to an international research collaboration?*

**A:** Desirable provisions include: clear specification of permitted uses and any re-use restrictions; explicit local co-authorship and capacity-building commitments (training, technology transfer); a commitment that a locally accessible copy of the dataset and any derived outputs remains with the Nigerian institution; clear ethical approval and consent documentation; and an agreed process for Nigerian researchers to review and approve any publication or further sharing of the data or findings.

***Q4.*** *Using the worked example in Section 16.5, explain why simply deleting the single at-risk record (rather than coarsening the quasi-identifiers) would itself introduce a new problem into the dataset.*

**A:** Deleting the at-risk record would silently reduce the sample size and could introduce a new, undocumented selection bias (systematically removing individuals with particular rare characteristics), whereas coarsening the quasi-identifiers preserves the informational content of the whole dataset for every individual (in a less granular but still valid form) rather than discarding a case entirely.

***Q5.*** *In Section 16.6, explain why the combination of lower AUC AND under-prediction of risk at primary health centres is a more serious equity concern than either finding alone.*

**A:** The combination is more serious than either alone because it means the model is not only less accurate (harder to trust its risk ranking) but systematically overconfident in a way that

would make a truly high-risk primary-health-centre patient appear lower-risk than they actually are, potentially delaying urgently needed referral or escalation of care precisely at the facilities with the fewest resources to catch this error through clinical judgement alone.

***PYTHON — supporting code***

```
auc_tertiary, auc_primary = 0.88, 0.74
underprediction_primary_pp = 12
print(f"AUC gap: {auc_tertiary - auc_primary:.2f}; "
    f"plus {underprediction_primary_pp}pp risk under-prediction at primary
facilities")
```

***Q6.*** *Using the R code in Section 16.5 as a guide, describe what would happen to the minimum group size (k) if 'ward' were replaced with the coarser variable 'LGA' in the grouping.*

**A:** Replacing 'ward' with the coarser variable 'LGA' would generally increase the minimum group size, since more individuals typically share the same LGA than share the same specific ward, meaning fewer (or no) combinations would fail a given k threshold — this illustrates the fundamental privacy/utility trade-off: coarser geography improves privacy protection but reduces the geographic precision available for spatial analysis (Chapter 14).

***PYTHON — supporting code***

```
import pandas as pd
df = pd.DataFrame({
   "age_band": ["0-9","0-9","10-19","10-19","20+"],
   "sex": ["F","F","M","F","M"],
   "ward": ["A","A","B","B","C"],
   "lga":  ["X","X","X","X","Y"],
})
k_by_ward = df.groupby(["age_band","sex","ward"]).size().min()
k_by_lga  = df.groupby(["age_band","sex","lga"]).size().min()
print(f"Min group size by ward: {k_by_ward}, by LGA (coarser): {k_by_lga}")
```

***Q7.*** *A funder asks you to justify, in two or three sentences, why 'the data has already had names removed' is not a sufficient answer to a privacy review question. Write that justification.*

**A:** A strong response notes that name removal alone does not address quasi-identifier-based re-identification risk (Section 16.2), that a formal privacy risk assessment (such as a k-anonymity check) is a recognised and expected step for responsible small-area or clinically-detailed health data release, and that failing to conduct one exposes both the research participants and the institution to real re-identification and reputational risk regardless of good intentions.

***Q8.*** *Using Section 16.8, explain why community trust, once damaged, could compromise future surveillance efforts beyond the immediate research project in question.*

**A:** Damaged community trust could compromise future surveillance efforts because communities that feel their health data was misused (for example, shared without consent, or used in a way that led to stigma or discrimination against community members) may become reluctant to participate honestly in future data collection, surveys, or outbreak investigations, undermining not just the original project but essential public health functions for years afterward.

***Q9.*** *Propose one concrete way a research team could meaningfully engage a community before collecting sensitive health data from its members.*

**A:** A concrete example: before beginning data collection, hold a meeting with recognised community leaders and representatives (traditional rulers, women's groups, patient advocacy groups relevant to the health topic) to explain in plain language what data will be collected, how it will be used and protected, and to invite questions and concerns, documenting the outcome of this engagement as part of the study's ethical approval materials.

***Q10.*** *Why might community engagement be especially important for research involving a stigmatised condition or a key population, compared to a general household health survey?*

**A:** Community engagement is especially important for stigmatised conditions or key populations because participants face amplified risks from any perceived data misuse or leak (social exclusion, discrimination, or even legal jeopardy in some contexts), and because such communities may already have justified historical distrust of research institutions, making proactive, respectful engagement essential to both ethical conduct and the practical success of the research.

***Q11.*** *In the Section 16.9 case study, explain why the fact that the original consent did not anticipate international data sharing matters, even though the new use is scientifically legitimate.*

**A:** The original consent's scope matters because informed consent is meant to reflect what participants (or their guardians) actually agreed to at the time; using data for a materially different, unanticipated purpose (international sharing for a new meta-analysis) without reconsidering consent risks violating the trust and autonomy the original consent process was meant to protect, regardless of the new use's scientific merit.

***Q12.*** *Propose one specific data transformation that could reduce re-identification risk for this dataset without eliminating its scientific value for the meta-analysis.*

**A:** A specific transformation: aggregating recruitment site information to a broader regional level (rather than sharing exact site identifiers) and widening age bands, both of which reduce the combination of quasi-identifiers available for potential re-identification while likely preserving enough scientific value for a larger multi-country meta-analysis focused on broader patterns rather than site-specific detail.

***Q13.*** *Why should the resolution of this kind of request generally involve the original institutional Health Research Ethics Committee, rather than being decided solely by the researchers holding the data?*

**A:** Involving the original Health Research Ethics Committee is important because they hold the institutional authority and contextual knowledge of the original consent process and participant population to make an informed judgment about whether the new use is adequately covered, or whether additional protections or a renewed consent process are required — a decision that should not rest solely with the researchers who happen to currently hold the data.

***Q14.*** *What would you want included in a data use and co-authorship agreement before agreeing to share this dataset internationally?*

**A:** A reasonable set of provisions: clear terms restricting the data's use to the specified meta-analysis purpose only; a requirement for any resulting publication to include appropriate co-authorship for the Nigerian research team; a data security and access-control commitment from the receiving consortium; and a stated process for the Nigerian institution to review findings before publication.

***Q15.*** *Explain why 'the request is well-intentioned and scientifically valuable' is necessary but not sufficient justification for approving a data sharing request like this one.*

**A:** Scientific value and legitimacy are necessary but not sufficient because research ethics review exists precisely to weigh scientific benefit against participant protection and equity concerns that are not automatically resolved just because a request comes from a well-regarded institution with good intentions — the specific risks (re-identification, consent scope, equitable benefit-sharing) must each be assessed on their own terms.

***Q16.*** *Using Section 16.10, explain why a study could have full research ethics approval but still fail to meet data protection compliance requirements.*

**A:** A study could have full ethical approval (correctly assessed as scientifically justified and ethically sound in its treatment of participants) while still processing the resulting personal data in a way that violates data protection requirements — for example, storing identifiable data insecurely, retaining it longer than lawfully permitted, or transferring it internationally without an adequate legal basis — since ethical approval addresses the RESEARCH DESIGN, not necessarily every subsequent detail of data handling and processing.

***Q17.*** *Under the Nigeria Data Protection Act 2023 framework referenced in Section 16.10, why would international data transfer be treated as a distinct issue requiring its own legal basis?*

**A:** International data transfer is treated as a distinct issue because moving personal data outside Nigeria's jurisdiction raises specific legal questions about which country's laws and protections will govern that data going forward, and about whether the receiving party or country provides an adequate level of protection — concerns that are separate from, and not automatically resolved by, the original research ethics approval for collecting the data within Nigeria.

***Q18.*** *Propose one practical step an institution could take to ensure both ethical approval and data protection compliance are checked, rather than assuming one covers the other.*

**A:** A practical step: establishing a formal institutional checklist or sign-off process requiring both a Health Research Ethics Committee approval AND a separate data protection/ compliance review (potentially involving a data protection officer or legal counsel) before any project involving personal

health data can proceed, rather than treating ethical approval as an implicit stand-in for both.

***Q19.*** *Explain the difference in focus between a Health Research Ethics Committee's review and a Data Protection Commission's concerns, in your own words.*

**A:** A Health Research Ethics Committee's review focuses on whether the RESEARCH itself is scientifically justified, uses appropriate methods, and adequately protects participants' rights, welfare, and informed consent; a Data Protection Commission's concerns instead focus specifically on the LAWFUL PROCESSING of personal data throughout its lifecycle — collection, storage, use, sharing, and disposal — regardless of the specific research purpose it serves.

***Q20.*** *Why is it important for a data governance process to treat these as 'two separate, both-necessary checks' rather than a single combined approval step?*

**A:** Treating these as two separate, both-necessary checks matters because relying on only one creates a genuine gap in protection — a scientifically and ethically sound study could still mishandle personal data in a way that causes real harm or legal exposure, and a data-protection-compliant data handling process does not by itself ensure the underlying research question and design are ethically justified in the first place.

***Q21.*** *Using Section 16.11, explain why remaining silent about discovered data misuse is described as an ethical problem, not just a matter of personal discomfort.*

**A:** Remaining silent about discovered data misuse is an ethical problem, not just a matter of personal discomfort, because doing so allows continued harm to research participants or communities whose data or trust has already been compromised, and it implicitly makes the silent observer complicit in that ongoing harm through inaction, even if they did not directly cause the original misuse.

***Q22.*** *Why is making an unverified public accusation also described as an inappropriate response, alongside remaining silent?*

**A:** An unverified public accusation is also inappropriate because it risks causing serious, potentially unjustified reputational harm to an individual or institution based on incomplete information, without first allowing the concern to be properly investigated through an appropriate, fair process — due process and verification matter even when a genuine ethical concern exists.

***Q23.*** *What is the recommended channel, according to Section 16.11, for reporting a genuine concern about data misuse?*

**A:** The recommended channel is a formal research integrity or ethics reporting mechanism maintained by the researcher's institution or the relevant international collaboration, specifically designed to receive, investigate, and appropriately act on such concerns through a fair and confidential process.

***Q24.*** *Why is it useful to know about this reporting pathway 'before you ever need it,' rather than only learning about it after discovering a problem?*

**A:** Knowing the reporting pathway in advance means a data scientist who discovers a genuine problem can act quickly and appropriately, without needing to first spend time figuring out where or how to raise the concern — a delay that could allow continued harm or make evidence harder to establish the longer it is discovered as a new problem in an unfamiliar process.

***Q25.*** *Propose a realistic (anonymised or hypothetical) scenario where you would feel obligated to use such a reporting channel, and explain why.*

**A:** A plausible scenario: discovering that a colleague has shared a dataset containing patient identifiers with an external party not covered by the original data use agreement, without the

knowledge or consent of the research participants or the institution's data governance office, which would appropriately be raised through the institution's formal research integrity reporting channel rather than addressed informally or ignored.

***Q26.*** *Summarise, in your own words, why this chapter treats data ethics review and formal data protection compliance as related but distinct processes (Section 16.10).*

**A:** Data ethics review and data protection compliance are treated as distinct because ethics review addresses whether a study is justified and respectful of participants' rights and welfare (a substantive, judgement-based process), while data protection compliance addresses specific technical and legal requirements for handling data securely (a more procedural, checklist-based process) — a project can satisfy one without automatically satisfying the other.

***Q27.*** *Explain the common thread linking the whistleblowing discussion (Section 16.11) and the reconsidered data-sharing request case study (Section 16.9).*

**A:** The common thread is that both whistleblowing and the reconsidered data-sharing request illustrate that good intentions alone do not guarantee ethical conduct — a data request framed as beneficial, or a colleague's data handling practice that seems locally reasonable, can still raise legitimate concerns that deserve scrutiny rather than automatic trust.

***Q28.*** *Why does this chapter argue that a 'well-intentioned' data request can still raise legitimate ethical concerns?*

**A:** A well-intentioned data request can still raise legitimate concerns because good intent does not eliminate downstream risks such as re-identification, scope creep beyond the originally consented purpose, or inadequate security at the receiving institution — the request's stated purpose does not, by itself, resolve these structural risks.

***Q29.*** *Propose a scenario from your own institution where distinguishing ethics review from data protection compliance (Section 16.10) would matter in practice.*

**A:** Answers will vary, but should identify a specific scenario (e.g., a dataset being shared for what is technically a new, unconsented secondary use) where an ethics committee's substantive judgement would be needed even if the data transfer itself technically satisfies a formal data protection checklist.

***Q30.*** *If you personally observed a colleague misusing sensitive health data, what would this chapter suggest as an appropriate first step?*

**A:** This chapter would suggest first raising the concern through the appropriate internal channel (a supervisor, an institutional ethics or data protection officer) rather than either ignoring it or taking unilateral public action, reflecting the general principle that data misuse concerns should be addressed through legitimate institutional accountability structures wherever they exist and are trusted to function.

***Q31.*** *Using Section 16.13, explain what is meant by 'undocumented ethical debt' in an inherited dataset.*

**A:** Undocumented ethical debt refers to a situation where a dataset's original ethical approval, participant consent language, and permitted-use restrictions are poorly documented, ambiguous, or entirely missing, creating uncertainty about whether a new, planned analysis is actually permitted under the data's original ethical foundation.

***Q32.*** *What specific steps should a researcher take before analysing a dataset originally collected by a different team years earlier?*

**A:** Before analysing an inherited dataset, a researcher should attempt to locate the original ethical approval documentation and consent forms/language, confirm the planned new analysis clearly falls within the originally approved and consented scope

of use, and, if this cannot be confirmed with reasonable confidence, consult the relevant institutional ethics committee for fresh guidance before proceeding.

***Q33.*** *Why is physical possession of a dataset not, by itself, sufficient justification for analysing and publishing from it?*

**A:** Physical possession of a dataset says nothing about whether the original research participants ever consented to the specific new use being planned; ethical and legal permission to analyse data derives from the original approval and consent process, not from simple access to the data files themselves.

***Q34.*** *Describe a plausible scenario where a planned new analysis of an inherited dataset would fall outside its original ethical approval's scope.*

**A:** A plausible scenario: a dataset originally collected and consented for a study specifically about maternal health outcomes being repurposed years later for an unrelated genetic or behavioural analysis that the original consent language never contemplated or explicitly permitted.

***Q35.*** *How does the concept of undocumented ethical debt relate to the data dictionary and metadata documentation practices introduced in Chapter 4?*

**A:** Both concepts share the principle that a dataset's full context — its variable definitions in the case of a data dictionary, or its ethical permissions and constraints in the case of ethical debt — must be explicitly documented and preserved over time, since undocumented context is easily lost as a dataset is passed between successive researchers or research teams.

## Chapter 17
## Capstone
### Case Studies and a Framework for Independent Practice

***Q1.*** *Choose a health data project from your own work. Apply the ten-point checklist in Section 17.7 and identify the two items you are currently weakest on.*

**A:** Answers will vary by individual, but a complete answer identifies at least two specific checklist items (e.g., 'documented reproducible cleaning' or 'checked subgroup performance') where the respondent's actual practice currently falls short, along with a brief, honest explanation of why.

***PYTHON — supporting code***

```
checklist = ["precise question", "known data source(s)", "reproducible cleaning",
        "EDA before modelling", "method matches goal", "out-of-sample
validation",
        "spatial/temporal structure handled", "assumptions & subgroup checks",
        "ethical approval", "reproducible end-to-end"]
weakest_two = checklist[6:8]  # example self-assessment
print("Weakest two items:", weakest_two)
```

***Q2.*** *For Case Study A, propose one additional analysis (drawing on any chapter of this book) that would strengthen the policy conclusions.*

**A:** Reasonable proposals include: a formal Moran's I spatial clustering test (Chapter 14) on the state-level TPR discrepancies to see if the HMIS-vs-register gap itself clusters geographically (perhaps concentrated in states with weaker reporting infrastructure), or a time series analysis (Chapter 10) of whether the discrepancy has been narrowing or widening over recent years as data systems have matured.

***Q3.*** *Design, in outline form only, a new capstone case study from a Nigerian health data source not covered in Sections 17.2-17.6, specifying which chapters' methods it would draw on.*

**A:** Answers will vary, but a complete outline names a Nigerian data source not used in Sections 17.2-17.6 (e.g., vital registration data, or an institutional dermatoglyphics/clinical cohort), states a specific research question, and explicitly names at least three chapters' methods the analysis would draw on (e.g., Chapter 4 cleaning, Chapter 9 survival analysis, Chapter 16 ethics review).

***Q4.*** *In Case Study E (Section 17.6), explain why recommending 'collect more data before deployment' is a more scientifically honest outcome than deploying the model with a known subgroup performance gap.*

**A:** This is more scientifically honest because deploying a model with a known, quantified performance gap at primary health centres (as revealed by the subgroup audit) would knowingly provide less reliable risk predictions to a setting the team has already identified as under-served by the model — recommending additional targeted data collection before deployment demonstrates the kind of systematic, evidence-based caution this book has argued for throughout, rather than prioritising a premature 'we built something' outcome over patient safety and equity.

***PYTHON — supporting code***

```
known_gap_pp = 12   # AUC gap between tertiary and primary facilities (Ch.16)
underprediction_pp = 12
print("Recommending more data collection rather than deploying, given a",
    known_gap_pp, "point AUC gap and", underprediction_pp, "pp risk under-
prediction")
```

***Q5.*** *For Case Study C, explain how you would decide whether the adolescent cohort's follow-up attrition is more consistent with MAR or MNAR (Chapter 4), and what analysis you would run to investigate this.*

**A:** One would compare characteristics of adolescents who dropped out at follow-up versus those who remained, focusing on whether dropout correlates with OBSERVED baseline variables (supporting MAR, e.g., dropout correlating with baseline food insecurity) versus showing evidence that dropout is additionally related to the (unobserved) follow-up symptom trajectory itself, such as through sensitivity analyses comparing

results under different plausible MNAR assumptions or by attempting targeted follow-up with a subsample of dropouts to directly observe their true outcome.

***Q6.*** *For Case Study D, explain why privacy safeguards (Chapter 16) are especially important for an RDS study of a key population, compared to a general population household survey.*

**A:** Privacy safeguards are especially important for an RDS study of a key population because participants may face significant social, legal, or safety risks (stigma, discrimination, or in some contexts legal jeopardy) if their identity or health status were disclosed, and because the peer-referral recruitment structure itself creates additional re-identification risk (participants may recognise their peers described in aggregated results) beyond what a general household survey would typically pose.

***Q7.*** *Write a short paragraph, in your own words, on which single habit from this entire book you consider most important to carry into your own future work, and why.*

**A:** Answers will vary by individual, but a strong response identifies a single, clearly articulated habit (e.g., 'always trace an apparent finding back to how the data was generated before trusting it,' or 'always validate a predictive model on data it did not see, respecting any natural grouping in the data') and explains, in the writer's own words, why that particular habit is the one they consider most valuable to carry forward.

***Q8.*** *Using Section 17.8, identify which of the four listed reasons projects stall you consider most common in your own experience or observation, and explain why.*

**A:** Answers will vary by individual experience, but a complete answer selects one of the four listed reasons and explains, with a specific example if possible, why it is the most commonly observed obstacle in the respondent's own context.

***PYTHON — supporting code***

```
stall_reasons = ["underestimated data access lead time", "skipped EDA",
                 "unnecessarily complex method", "no reproducible project structure"]
# Self-audit: rank which of these has caused delay in your own recent project
for i, reason in enumerate(stall_reasons, 1):
    print(i, reason)
```

***Q9.*** *Explain how skipping Chapter 5's exploratory analysis stage could lead to a costly restart later in a project.*

**A:** Skipping exploratory analysis means data quality problems (implausible values, unexpected missingness patterns, coding inconsistencies) that EDA would typically surface early are instead discovered much later — often only when a formal model produces a nonsensical result — by which point substantial modelling work may need to be redone from scratch once the underlying data issue is identified and fixed.

***Q10.*** *Propose one practical habit (from anywhere in this book) that would directly address one of the four stall points listed in Section 17.8.*

**A:** Answers will vary, but a strong response names a specific habit from any chapter (e.g., 'always build a data provenance log before starting analysis,' from Chapter 1) and explains concretely how consistently applying it would have prevented or mitigated one of the four listed stall points.

***Q11.*** *In the Section 17.9 case study, explain why the pre-existing difference in baseline ANC attendance between programme and comparison LGAs was a warning sign that needed to be addressed before drawing conclusions.*

**A:** The pre-existing difference in baseline ANC attendance was a warning sign because it suggested the programme-deployment and comparison LGAs were not truly comparable from the start, meaning any later difference between the groups could reflect this pre-existing gap (and whatever underlying factors caused

it) rather than a genuine effect of the programme itself — a classic confounding-by-selection concern.

***Q12.*** *Why was a difference-in-differences approach more appropriate here than a simple before-after or treatment-vs-comparison comparison alone?*

**A:** A difference-in-differences approach was more appropriate because it explicitly accounts for each group's own pre-existing trajectory, comparing the CHANGE over time in each group rather than simply comparing their levels at one point in time — this specifically corrects for a stable, pre-existing gap between the groups that a simple before-after or treatment-vs-comparison snapshot comparison would not.

***Q13.*** *Explain why the research team's supervisory committee judged the honestly reported null result to be stronger work than an uncritical positive finding would have been.*

**A:** The committee judged the null result stronger work because it demonstrated the full methodological rigour this book has argued for — careful data provenance, exploratory investigation of a confounding risk, and an appropriately adjusted analytic method — applied faithfully and reported honestly even when the outcome was not the one the team had hoped for, which is a more reliable demonstration of genuine analytical competence than an uncritically confirmed positive finding.

***Q14.*** *Propose one additional outcome (beyond early ANC attendance) that this programme's benefits might plausibly show up in, and how you would test for it.*

**A:** A plausible additional outcome: attendance at later ANC visits (e.g., the fourth or subsequent visit) or postnatal care uptake, both of which could plausibly be more directly influenced by ongoing community health worker contact than a single early-visit indicator; testing this would require the same difference-in-differences approach applied to these alternative outcome variables.

***Q15.*** *What is the risk, for the research team's own credibility, of searching for an alternative specification that would produce the originally hoped-for positive result?*

**A:** The risk is that searching for an alternative specification specifically because it would produce a hoped-for result — rather than because it is independently well-justified — constitutes exactly the kind of undisclosed multiple-comparisons/p-hacking risk discussed in Chapter 7, which, if later uncovered by a reviewer or replication attempt, could seriously damage the research team's credibility and the credibility of the broader finding.

***Q16.*** *Using Section 17.10, explain why a co-author who can read a well-documented, version-controlled script can contribute more effectively than one who only sees a final results table.*

**A:** A co-author who can read a well-documented, version-controlled script gains direct, verifiable insight into exactly what analytic choices were made and why, allowing them to spot potential issues, suggest specific improvements, or confirm their confidence in the approach — a co-author who only sees a final results table must either trust the results blindly or ask the original analyst to explain every detail verbally, which is far less efficient and more error-prone.

***Q17.*** *Propose one specific agreement a research team should make at the start of a collaboration to avoid the 'two different versions of the cleaned dataset' problem described in Section 17.10.*

**A:** A specific agreement: designating a single shared, version-controlled repository (Chapter 6) as the definitive location for all project code and data, with an explicit rule that any team member's local copy is considered provisional until changes are committed and shared back to this single shared location, preventing parallel, undocumented divergence.

***Q18.*** *Why does establishing pipeline responsibilities early (Section 1.2) help prevent friction later in a collaborative project?*

**A:** Establishing pipeline responsibilities early clarifies exactly who is expected to produce which artefact at which pipeline stage (Section 1.2) — for example, who owns the cleaning script versus who owns the final modelling script — reducing the chance that two team members unknowingly duplicate or silently diverge in producing what should be a single, shared version of a given pipeline stage's output.

***Q19.*** *Describe a situation (real or plausible) where a lack of shared, version-controlled workflow caused confusion or wasted effort in a team project.*

**A:** Answers will vary, but a plausible example: two team members independently 'fix' the same data quality issue in their own separate copies of a cleaning script using two different exclusion criteria, and only discover the discrepancy much later when their reported sample sizes in a shared manuscript draft do not match.

***Q20.*** *How would you raise this issue diplomatically with a co-author or supervisor who is used to working without version control?*

**A:** A diplomatic approach would frame the request around a shared, non-personal goal — for example, 'I think a shared repository would help us stay in sync and avoid us accidentally working from different versions of the cleaned dataset' — offering to set up and demonstrate the workflow rather than simply asserting that the collaborator's current approach is inadequate.

***Q21.*** *Using Section 17.11, rewrite the example phrase 'the Poisson regression coefficient for LGA was significant at $p<0.05$' into a plain-language, action-oriented statement suitable for a non-technical LGA health officer.*

**A:** A plain-language rewrite: 'This LGA's malaria testing shows a pattern that's different from most other LGAs and worth a closer look before the rainy season starts,' replacing the technical statistical language with a direct, action-relevant statement while preserving the finding's substance.

***Q22.*** *Why is being prepared to answer 'so what should we do about it' described as a natural and expected follow-up question?*

**A:** This question is natural and expected because a non-technical decision-maker's primary interest in any presented finding is what action it implies they should take, not the statistical method that produced it — anticipating and preparing a clear, specific answer to this question is essential to making the presented finding genuinely useful rather than merely informative.

***Q23.*** *Explain the risk of using too much technical hedging language when communicating a finding's limitations to a non-technical audience.*

**A:** Excessive technical hedging (e.g., extensive caveats about confidence intervals, model assumptions, and alternative specifications, presented in full statistical detail) risks obscuring the finding's core, actionable message entirely, potentially causing a non-technical audience to either disengage from an overly complex presentation or misunderstand which parts of the finding are solid versus genuinely uncertain.

***Q24.*** *Propose one visualisation choice (from Chapter 5) that would likely communicate a key finding more effectively to a community leader than a dense data table.*

**A:** A simple, well-labelled bar chart or trend line showing the key comparison directly (e.g., malaria case counts before and after an intervention, or coverage rates across LGAs) would likely communicate a finding far more effectively to a

community leader than a dense table of regression coefficients and p-values, which requires statistical training to interpret correctly.

***Q25.*** *Why does this book argue that this communication skill develops mainly through direct practice, rather than through reading about it alone?*

**A:** This skill develops mainly through direct practice because effectively gauging a specific non-technical audience's background knowledge, concerns, and likely follow-up questions in real time is a practical, situational skill that general principles alone (however well internalised) cannot fully substitute for — actual experience presenting to real audiences, and learning from their actual reactions and questions, is what builds this skill most effectively.

***Q26.*** *Summarise, in your own words, why this chapter's checklist (Section 17.7) is presented as a practical summary of the entire book rather than a new set of ideas.*

**A:** The checklist is presented as a summary rather than new material because every one of its ten items is drawn directly from a specific chapter's core argument — the checklist's value lies in consolidating the whole book's lessons into a single, practically usable reference, not in introducing any genuinely new concept.

***Q27.*** *Explain the common thread linking the contradicted-hypothesis case study (Section 17.9) and the non-technical presentation discussion (Section 17.11).*

**A:** The common thread is that both examples highlight that good data science practice extends beyond the technical analysis itself into how results are handled and communicated — deciding how to report a result that contradicts one's own hypothesis, and adapting a technical finding for a non-technical audience, both require judgement beyond correct statistical execution.

***Q28.*** *Why does this chapter emphasise working constructively with co-authors and supervisors (Section 17.10) as a data science skill in its own right?*

**A:** Working constructively with co-authors and supervisors is treated as a genuine data science skill because most real research and applied analysis happens collaboratively, and an analyst's technical work only creates value if it can be integrated into, and trusted by, a broader team — a skill distinct from, but just as necessary as, technical competence.

***Q29.*** *Propose a hypothetical capstone project from your own field, and describe how you would present its key finding to a non-technical audience (Section 17.11).*

**A:** Answers will vary, but should describe a specific hypothetical project and a plausible, plain-language way of presenting its key finding without technical jargon, appropriate for a non-technical stakeholder audience.

***Q30.*** *If your own capstone analysis contradicted your initial hypothesis, as in Section 17.9's case study, how would you decide what to report?*

**A:** A reasonable response emphasises reporting the contradictory finding honestly and completely, following this book's Chapter 5 caution against treating any single analysis as automatically confirmatory, and framing the unexpected result as a genuine, interesting finding in its own right rather than something to downplay or explain away.

***Q31.*** *Using Section 17.13, list the four key elements a strong methods section should include, according to this chapter.*

**A:** The four key elements: (1) the data source and study design, (2) cleaning and exclusion decisions with their rationale, (3) the exact statistical or machine learning method used, with enough detail for another researcher to replicate it, and (4) how missing data, multiple comparisons, or model validation were specifically handled.

***Q32.*** *Why is a methods section described as a 'prose summary' of the same decisions documented throughout this book's mini-projects?*

**A:** A methods section is described this way because every one of the decisions it reports (why data was excluded, which model was chosen, how missingness was handled) is exactly the kind of decision this book has repeatedly asked readers to make explicitly and document throughout their mini-project write-ups — the methods section is simply that same documentation, condensed and reformatted for formal publication.

***Q33.*** *Choose one of your own mini-project write-ups from this book and draft a two-to-three sentence methods section based on it.*

**A:** Answers will vary depending on the reader's own completed mini-project, but a strong response is a concise, factual paragraph naming the dataset, the key cleaning/exclusion decisions, the specific statistical method applied, and how any missing data or validation step was handled — in formal, past-tense scientific prose.

***Q34.*** *Why is stating how missing data and multiple comparisons were handled considered essential for a complete methods section?*

**A:** Stating how missing data and multiple comparisons were handled is essential because these decisions can materially affect a study's conclusions and its risk of false-positive findings, and a reader (or reviewer) needs this information to correctly judge how much confidence to place in the reported results.

***Q35.*** *What is the risk of omitting the rationale for a cleaning or exclusion decision from a published methods section?*

**A:** Omitting the rationale for a cleaning or exclusion decision means a reader cannot judge whether that decision was reasonable or could have introduced bias, and it also means the

analysis cannot be faithfully replicated by another researcher attempting to reproduce the study using only the published methods description.

## Chapter 18

## Introduction to Python for Disease Epidemiology

***Q1.*** *Explain, using the shape of an epidemic curve, how you would distinguish a point-source outbreak from one driven by ongoing person-to-person transmission.*

**A:** A point-source outbreak typically shows a single sharp peak, roughly one incubation period wide, followed by a decline, consistent with many people being exposed to a common source (e.g., contaminated water) at roughly the same time; ongoing person-to-person transmission instead typically shows a series of successive peaks spaced roughly one serial interval apart, or a more sustained, slowly-declining tail, reflecting successive generations of transmission rather than a single shared exposure event.

### *PYTHON — supporting code*

```
import matplotlib.pyplot as plt
point_source_pattern = [2, 8, 25, 40, 30, 15, 6, 2]      # single sharp peak
person_to_person_pattern = [2, 5, 4, 9, 8, 14, 12, 18]   # successive rising peaks
fig, ax = plt.subplots(1, 2, figsize=(8, 3))
ax[0].bar(range(8), point_source_pattern); ax[0].set_title("Point-source")
ax[1].bar(range(8), person_to_person_pattern); ax[1].set_title("Person-to-person")
```

***Q2.*** *A line list has 50 confirmed cases, of which 12 have died, 30 have recovered, and 8 are still hospitalised with an unknown outcome. Calculate both the naive CFR (using all 50 cases as the denominator) and the corrected CFR (using only resolved cases), and explain which is more appropriate to report while the outbreak is still active.*

**A:** Naive CFR = 12/50 = 24.0%. Corrected CFR (resolved cases only, 12 deaths out of 12+30=42 resolved) = 12/42 H" 28.6%. The corrected figure is more appropriate to report while the outbreak is active, because the naive calculation includes the 8 still-hospitalised patients with unknown outcomes in its denominator, which understates the true proportion of resolved cases that end in death (some of those 8 unresolved cases may yet die, but they should not simply be counted as 'survivors' by omission from the numerator while still being counted in the denominator).

***PYTHON — supporting code***

```
died, recovered, unresolved, total = 12, 30, 8, 50
naive_cfr = died / total
corrected_cfr = died / (died + recovered)
print(f"Naive CFR: {naive_cfr:.1%}, Corrected CFR: {corrected_cfr:.1%}")
# Naive CFR: 24.0%, Corrected CFR: 28.6%
```

***Q3.*** *Explain, in your own words, the difference between R0 (Chapter 13) and Rt (Section 18.6), and why Rt is generally more useful for monitoring an outbreak in real time.*

**A:** R0 is a single, fixed summary describing the expected number of secondary infections in a fully susceptible population at the START of an outbreak, under the assumed baseline transmission conditions; Rt is a time-varying quantity that reflects the current, actual transmission rate at any given point in the outbreak, incorporating the real effects of accumulating immunity, behaviour change, and interventions already in place — Rt is more useful for real-time monitoring because it directly answers 'is the outbreak currently growing or shrinking right now,' which R0 alone cannot tell you once an intervention or population immunity has begun to change transmission.

***Q4.*** *Using the worked example in Section 18.10, explain why finding that cases cluster significantly around two specific*

*wards supports the point-source hypothesis rather than the person-to-person transmission hypothesis.*

**A:** Finding statistically significant spatial clustering specifically around the two wards nearest the implicated water point supports the point-source hypothesis because sustained person-to-person transmission would typically be expected to spread more broadly and progressively across the wider area over 90 days as chains of transmission extend outward, whereas a tight, geographically localised cluster concentrated near a specific shared exposure point is a hallmark signature of a common-source outbreak.

***Q5.*** *Using the code in Section 18.5, recompute the corrected CFR if 4 more of the 'unresolved' cases in the example were to later die, and comment on how much the CFR estimate could still change as an outbreak resolves.*

**A:** If 4 more of the 8 unresolved cases die, resolved deaths become 12+4=16 out of a new resolved total of 42+4=46 (using the recomputed number from the original scenario in Exercise 2), giving a corrected CFR of 16/46 H" 34.8% — a substantial further increase from the original 28.6% estimate, illustrating how much a CFR estimate calculated mid-outbreak can still change as more outcomes resolve, which is exactly why it should always be reported with the caveat that it is provisional while unresolved cases remain.

***PYTHON — supporting code***

```
died, recovered, unresolved = 12, 30, 8
extra_deaths = 4
new_died = died + extra_deaths
new_resolved = died + recovered + extra_deaths
new_cfr = new_died / new_resolved
print(f"Updated corrected CFR: {new_cfr:.1%}")
# Updated corrected CFR: 34.8%
```

***Q6.*** *In Section 18.9, explain why a logistic regression's coefficients being reported (rather than only its AUC) is useful for a field epidemiology team trying to understand which symptoms most elevate risk of a severe outcome.*

**A:** Reporting the coefficients (not just the AUC) tells the field team which specific symptoms or characteristics are most strongly associated with elevated severity risk (for example, if 'altered_consciousness' has a much larger positive coefficient than 'age'), which can directly inform triage protocols and clinical training on which presenting signs should trigger the most urgent escalation, information the overall AUC alone (a single number describing general discriminative ability) does not provide.

***Q7.*** *Using Section 18.6's Rt approach, explain what a sustained Rt estimate above 1 for several consecutive time windows would suggest to an outbreak response team, and what action this might prompt.*

**A:** A sustained Rt estimate above 1 across several consecutive time windows would suggest the outbreak is continuing to grow rather than being brought under control, and would likely prompt the response team to intensify or reconsider their current interventions (e.g., strengthen case isolation, expand contact tracing, investigate whether a suspected common source has actually been fully remediated) rather than beginning to scale down the response.

***Q8.*** *Using Section 18.11, explain one advantage of converting a personal analysis script into a shareable command-line tool for a field surveillance team.*

**A:** A command-line tool means a field surveillance team member does not need to understand or edit the underlying Python code at all — they can run a single, well-documented command with a new data file and receive the same standardised outputs (epicurve, attack rate, Rt estimate) every time, without

needing programming expertise or risking accidentally introducing an error into the analysis code itself.

***Q9.*** *Why might an interactive dashboard (e.g., built with Streamlit) be more useful than a script for a non-programmer team member monitoring an ongoing outbreak?*

**A:** An interactive dashboard is more useful for a non-programmer because it provides an immediate, visual, point-and-click interface (uploading a file, viewing updated charts) rather than requiring any command-line interaction at all, making it accessible to team members with no technical background whatsoever, which is often the reality for front-line surveillance staff.

***Q10.*** *What risk arises if institutional infrastructure (like the tool described in Section 18.11) depends entirely on one analyst's continued involvement, and how might a team mitigate this risk?*

**A:** If institutional infrastructure depends entirely on one analyst, the tool risks becoming unmaintained or unusable the moment that analyst leaves, changes roles, or is simply unavailable when an urgent fix or update is needed; a team can mitigate this by documenting the tool's code and design decisions clearly, training at least one additional team member to maintain it, and using version control (Chapter 6) so the tool's history and logic remain accessible to whoever needs to pick it up next.

***Q11.*** *In the Section 18.12 case study, explain how a day/month transposition error could occur even when data entry staff are experienced and careful.*

**A:** A day/month transposition error can occur even with careful, experienced staff because different data entry systems, training materials, and personal habits sometimes default to different date formats (DD/MM/YYYY versus MM/DD/YYYY), and a date such as 3 June (03/06) versus 6 March (06/03) is genuinely

ambiguous without knowing which format convention was intended, making this a systematic risk built into the data entry process itself rather than a simple one-off mistake.

***Q12.*** *Why did correcting the three flagged records (rather than excluding them) matter for the outbreak's exposure-window investigation specifically?*

**A:** Correcting the three records mattered because two of them fell within the outbreak's crucial first week, and including them (with corrected dates) moved the estimated true earliest case four days earlier than the uncorrected data suggested — directly shifting the plausible exposure window investigators used to identify which water sources and community events might be linked to the outbreak's origin.

***Q13.*** *Propose one systems-level change (e.g., to the data entry form or software) that could prevent this specific type of date error from occurring in future outbreaks.*

**A:** A systems-level change could be standardising all data entry forms and software to a single, unambiguous date format (for example, always displaying and requiring entry as DD-Mon-YYYY, such as 03-Jun-2024, which cannot be confused between day and month), removing the ambiguity that allows this specific transposition error to occur in the first place.

***Q14.*** *Why is it significant that the validation check in Section 18.3, introduced early in this chapter, was what surfaced this error rather than a later, more complex analysis step?*

**A:** It is significant because it demonstrates that a simple, early-stage validation check (introduced as a basic data-quality step in Section 18.3) can have consequences for a field investigation's most substantive conclusions, reinforcing that data-quality discipline is not merely a preliminary formality to complete before the 'real' analysis begins, but an integral part of the analysis that can materially change its findings.

***Q15.*** *Write a two-sentence closing reflection, in your own words, connecting this final case study back to the very first chapter's argument about data provenance.*

**A:** A reasonable closing reflection: this book opened by arguing that understanding where data comes from matters more than the sophistication of any single statistical technique, and this final case study shows that argument holding true from the very first data-entry step through to the final epidemiological conclusion — the discipline of questioning and verifying data, not any particular formula or algorithm, is this book's most durable lesson.

***Q16.*** *Using Section 18.13, explain why Python's advantage is most decisive when a pipeline must be integrated into a broader software system.*

**A:** Python's advantage is most decisive here because its general-purpose programming ecosystem (web frameworks, API libraries, broader software engineering tooling) is considerably more mature and widely supported than R's for building and maintaining production software systems, making it the more natural choice when an analytic pipeline needs to become a durable piece of institutional software infrastructure (as in Section 18.11) rather than remaining a standalone analysis.

***Q17.*** *For which specific chapters or tasks in this book does Section 18.13 suggest R's ecosystem remains the more natural first choice?*

**A:** Section 18.13 suggests R's ecosystem remains the more natural first choice for classical statistical inference, complex survey design analysis (Chapter 7), and survival analysis (Chapter 9), where R's longer-established statistical packages and conventions are particularly mature and widely used within the biostatistics community.

***Q18.*** *Explain why this book takes the position that 'the right tool depends on the task,' rather than recommending one language over the other in general.*

**A:** This book takes this position because it has consistently argued (from Chapter 6 onward) that tool choice should be driven by the specific requirements of the task at hand — data structure, deployment context, team expertise, and ecosystem maturity for that particular kind of work — rather than by a fixed loyalty to one programming language regardless of fit, a principle applied consistently across every method-selection decision in this book.

***Q19.*** *Propose a Nigerian health data science task not explicitly mentioned in Section 18.13 where you would lean toward Python, and justify your choice.*

**A:** A plausible example: building an automated, continuously-updating outbreak surveillance dashboard that ingests new case reports and displays live-updating maps and charts to a broad team — Python's web-framework and interactive-dashboard libraries (such as Streamlit, mentioned in Section 18.11) make this kind of production deployment considerably more straightforward than the equivalent in R.

***Q20.*** *Propose a Nigerian health data science task where you would lean toward R instead, and justify your choice, drawing on any relevant earlier chapter.*

**A:** A plausible example: designing and analysing a complex, multi-stage stratified household survey with intricate sampling weights and clustering (as in an NDHS-style analysis, Chapter 7) — R's 'survey' package and long-established conventions for this specific kind of design-based inference are generally more mature, more widely taught in biostatistics training, and more directly applicable than the comparatively newer and less standardised Python survey-analysis tooling.

***Q21.*** *Using Section 18.14, explain in your own words why this book argues fluency in a specific tool is less durable than the underlying habits of mind it has tried to teach.*

**A:** Fluency in a specific tool (a particular function, package, or even programming language) is less durable because software and libraries continually change, are superseded, or fall out of use over a career, while the underlying habits of disciplined thinking — questioning data provenance, distinguishing significance from importance, validating out-of-sample, communicating honestly — remain valuable and applicable regardless of which specific tools happen to be in use at any given time.

***Q22.*** *Which of the four habits listed in Section 18.14 (provenance, significance vs. importance, out-of-sample validation, honest communication) do you personally find most challenging to apply consistently, and why?*

**A:** Answers will vary by individual, but a thoughtful response honestly identifies one of the four habits (e.g., resisting the temptation to report a striking but unconfirmed EDA finding as a finding, per Chapter 5) and explains a specific, personal reason why that particular discipline is challenging to maintain consistently under real-world time or professional pressure.

***Q23.*** *Why does the chapter emphasise that R and Python are 'not fundamentally different in their underlying statistical and epidemiological logic,' despite their different syntax?*

**A:** The chapter emphasises this because a reader should understand that switching between R and Python (or eventually to some future tool) does not require re-learning the fundamental statistical and epidemiological reasoning this book has taught — only the specific syntax for expressing that same reasoning changes, meaning the deeper skills developed throughout this book transfer directly to whatever tool a reader uses next.

***Q24.*** *Explain how the final chapter's closing argument connects back specifically to the four-question checklist introduced in Section 1.9.*

**A:** The connection is that the final chapter's four listed habits are, in essence, an expanded, generalised version of the four-question provenance checklist introduced at the very start of the book in Section 1.9 — both emphasise pausing to interrogate a dataset's origins and limitations before drawing or communicating a conclusion from it, applied consistently from the book's first chapter to its very last.

***Q25.*** *Write a short, personal closing reflection (3-5 sentences) on which single idea from this entire book you intend to apply first in your own work, and why.*

**A:** Answers will vary by individual, but a genuine closing reflection should identify one specific, concrete idea from the book (not a vague generality) and briefly explain a plausible first situation in the reader's own work where they intend to apply it, demonstrating genuine engagement with the book's central arguments rather than a generic closing statement.

***Q26.*** *Summarise, in your own words, why this chapter closes the book by returning explicitly to Chapter 1's central argument (Section 18.14).*

**A:** The book closes by returning to Chapter 1 because the core argument — that trustworthy analysis depends on provenance-aware, validated, honestly communicated work, not on the sophistication of any particular tool or technique — is the single thread that has run through all eighteen chapters, and this final reflection makes that unifying argument explicit.

***Q27.*** *Explain the common thread linking the self-correcting pipeline case study (Section 18.12) and the discussion of when Python's advantages are most decisive (Section 18.13).*

**A:** The common thread is that both the self-correcting pipeline case study and the discussion of Python's comparative strengths illustrate that good tooling and good analytic practice reinforce each other — a well-built validation check caught a real data problem precisely because the pipeline was designed with the same disciplined, provenance-aware mindset this book has argued for throughout.

***Q28.*** *Why does this chapter argue that a well-built validation check (as in Section 18.12) is as valuable as the analysis itself?*

**A:** A well-built validation check is as valuable as the analysis itself because an analysis built on undetected bad data can produce a confidently wrong conclusion just as easily as, or more easily than, no analysis at all — the validation check is what makes the downstream analysis trustworthy in the first place.

***Q29.*** *Propose one scenario from your own field where Python's specific strengths (Section 18.13) would make it clearly preferable to R, and one where the reverse would be true.*

**A:** Answers will vary, but should identify a genuine relative strength of Python (e.g., its dominance in deep learning and general-purpose software integration) and a genuine relative strength of R (e.g., its long-standing depth in classical biostatistical and survival analysis packages) matched to specific plausible scenarios.

***Q30.*** *Having completed this book, what single practice do you now consider most essential to carry into your own future outbreak or health data analysis work?*

**A:** Answers will vary by individual, but a strong response names one specific, concrete practice (for example, always documenting data provenance before analysis, or always validating a model on data it has not seen) and explains, in the reader's own words, why they consider it the single most valuable habit from the entire book.

***Q31.*** *Using Section 18.16, explain why fitting the same model specification in R and Python on identical data should produce numerically similar results.*

**A:** Fitting the same model specification (the same predictors, the same functional form, the same outcome) on identical data should produce numerically similar results because logistic regression, like most classical statistical models, has a mathematically well-defined optimal solution (via maximum likelihood estimation) that does not depend on which software implementation is used to find it, provided both implementations correctly solve the same underlying optimisation problem.

***Q32.*** *If R and Python produced meaningfully different coefficients for the same model and data, what would you check first?*

**A:** If R and Python produced meaningfully different coefficients for the same nominal model and data, the first things to check would be whether the reference/baseline category for any categorical predictor was set identically in both languages, whether any predictor was scaled or transformed differently, and whether both languages' default convergence criteria and optimisation methods were configured comparably.

***Q33.*** *Why is this cross-language check a useful practical habit for an analyst who regularly works in both R and Python?*

**A:** This cross-language check is a useful habit because it provides an independent verification that a model has been correctly translated and specified in each language — a discrepancy immediately signals a likely specification error rather than requiring the analyst to simply trust that the translation between languages was done correctly.

***Q34.*** *Name one common source of a legitimate (not erroneous) discrepancy between two languages' implementations of 'the same' model.*

**A:** A common, legitimate source of discrepancy is a difference in the numerical optimisation algorithm's default convergence tolerance or maximum iteration count between the two languages' implementations, which can produce very slightly different (though practically negligible) coefficient estimates even when both models are correctly and equivalently specified.

***Q35.*** *Reflecting on the whole of Chapter 18, what would you say is the single most transferable skill from this chapter to a context outside outbreak investigation specifically?*

**A:** Answers will vary by individual, but a strong response identifies a specific, broadly transferable skill emphasised throughout the chapter — such as validating data before analysis, or always benchmarking a model or forecast against a simple baseline — and explains why it applies just as well to health data science tasks outside outbreak investigation specifically.