

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 = $(2.00 - 1.43) / 1.43 \times 100$ 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 = $0.95 \times 400 = 380$; false positives = $0.10 \times 600 = 60$; PPV = $380 / (380 + 60) = 380 / 440 \approx 86.4\%$. 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 = $(1 - \text{sensitivity}) \times \text{number infected} = 0.05 \times 400 = 20$ 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: $PPV = (0.85 \times 0.05) / [(0.85 \times 0.05) + (0.02 \times 0.95)] = 0.0425 / (0.0425 + 0.019) = 0.0425/0.0615$

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 $P(\text{Disease} \mid \text{Positive test})$ and $P(\text{Positive test} \mid \text{Disease})$, and why confusing the two is a common and consequential error in interpreting diagnostic tests.*

A: $P(\text{Disease} \mid \text{Positive test})$ is the probability a patient actually has the disease given they tested positive (what a clinician wants to know — the positive predictive value); $P(\text{Positive test} \mid \text{Disease})$ 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 = $0.90 \times 0.02 = 0.018$; false positives = $0.05 \times 0.98 = 0.049$; $PPV = 0.018 / (0.018 + 0.049) = 26.9\%$ — 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 LGAX 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 $n_1=60$, $n_2=40$ (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 $600/3.1 \approx 194$ 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 $d = \text{mean difference} / \text{pooled SD} = 140 / 425 \approx 0.33$, 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 (λ) well beyond ‘ $\lambda_{\min}$ ’.*

A: Increasing the LASSO penalty (λ) 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 R_0 , 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 R_0 (since R_0 is proportional to the transmission rate divided by the recovery rate); if the intervention is strong enough to push R_0 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 R_0 achieved by the intervention scenario, and comment on whether this alone would be enough to push R_0 below 1.*

A: R_0 reduction: baseline $R_0 = 0.30/0.10 = 3.0$; intervention $R_0 = 0.18/0.10 = 1.8$. Percentage reduction = $(3.0 - 1.8)/3.0 = 40.0\%$. Although this is a substantial reduction, $R_0 = 1.8$ 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 R_0 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 $t_0 = 6.2$ tells a health programme manager about the timing of the outbreak’s fastest growth phase.*

A: $t_0 = 6.2$ 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 $(1 - \text{vaccination_coverage})$ 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 R_0 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 R_0 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 ($\text{beta} \times S \times I / N$, 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 $R_0 = 5$, using the formula in Section 13.10.*

A: Herd immunity threshold = $1 - 1/R_0 = 1 - 1/5 = 1 - 0.2 = 0.8$, meaning approximately 80% of the population must be immune to interrupt sustained transmission for a disease with $R_0 = 5$.

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 $R_0 = 15$ for measles associated with such a high (93%) herd immunity threshold, compared to a disease with a lower R_0 ?*

A: A higher R_0 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 ($1 - 1/R_0$) mathematically increases as R_0 increases, which is why measles, with its very high R_0 , 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) = $12/42 = 28.6\%$. 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 R_0 (Chapter 13) and R_t (Section 18.6), and why R_t is generally more useful for monitoring an outbreak in real time.*

A: R_0 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; R_t 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 — R_t is more useful for real-time monitoring because it directly answers ‘is the outbreak currently growing or shrinking right now,’ which R_0 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 R_t approach, explain what a sustained R_t estimate above 1 for several consecutive time windows would suggest to an outbreak response team, and what action this might prompt.*

A: A sustained R_t 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, R_t 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.