
Data Science for the Health Sciences

Data Science for the Health Sciences

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

Oluwafemi Samuel Oyamakin

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

University Post Office
Box 20984
Ibadan
Oyo State, Nigeria

7 Ikioda Street
Jattu-Uzairue
Edo State, Nigeria

© Oluwafemi Samuel Oyamakin

First Edition: 2026

ISBN: 978-978-035

All Rights Reserved

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

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

Dedication

Acknowledgements

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

Preface

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

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

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

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

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

How to Use This Book

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

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

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

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

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

Notation and Abbreviations

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

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

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

Contents

Dedication	v
Acknowledgements	vi
Preface	vii
Notation and Abbreviation	x

PART I: Foundations

Chapter 1: Introduction to Data Science for the Health Sciences	3
Chapter 2: The Nigerian Health Data Ecosystem	21
Chapter 3: Foundations of Probability and Statistical Thinking for Health Data	37

PART II: Data Wrangling and Exploratory Analysis

Chapter 4: Data Management: Types, Sources, Collection, and Cleaning	55
Chapter 5: Exploratory Data Analysis and Visualization	71
Chapter 6: Programming Foundations in R and Python for Health Data Science	86

PART III: Statistical Inference and Modelling

Chapter 7:	Statistical Inference and Hypothesis Testing in Health Research	103
Chapter 8:	Regression Modelling: Linear, Logistic, and Beyond	118
Chapter 9:	Survival Analysis and Time-to-Event Data	134
Chapter 10:	Time Series Analysis and Disease Surveillance Forecasting	149

PART IV: Machine Learning and Advanced Methods

Chapter 11:	Machine Learning Fundamentals for Health Prediction	166
Chapter 12:	Unsupervised Learning: Clustering and Dimensionality Reduction	181
Chapter 13:	Mathematical Modelling of Infectious Diseases	195
Chapter 14:	Spatial Epidemiology and Disease Mapping	211

PART V: Big Data, Ethics, and Practice

Chapter 15:	Big Data, AI, and Deep Learning in Health Systems	227
Chapter 16:	Data Ethics, Governance, and the Future of Health Data Science in Nigeria	242
Chapter 17:	Capstone: Case Studies and a Framework for Independent Practice	256

PART VI: Python for Disease Epidemiology

Chapter 18: Introduction to Python for Disease Epidemiology	273
Glossary of Key Terms	293
Selected References and Further Reading	298
Appendix: Answers to End-of-Chapter Exercises (All 18 Chapters)	301

Part 1

Foundations

