Applied Statistical Methods for Health Sciences Research

A graduate textbook

The rgtlab Curriculum Project

2026-04-29

GRADUATE BIOSTATISTICS SERIES

rgtlab Curriculum Project

Applied Statistical Methods for Health Sciences Research

A graduate textbook

First Edition · 2026

rgtlab

Welcome

This is the online version of **Applied Statistical Methods for Health Sciences Research** by The rgtlab Curriculum Project, a graduate textbook.

The book covers the applied methodological core of an MS-level biostatistics curriculum: estimands and study design, causal inference for observational data, mediation, longitudinal and survival analysis at applied depth, clinical trial design and analysis, missing-data methodology, meta-analysis, and advanced categorical-data methods. It is positioned between the workflow-focused *Practicum* and the computing-focused *SCAI* volumes on the methods axis.

The book sits in a five-volume graduate sequence:

- *R for Biostatistics: A One-Week Boot Camp* — pre-program preparation.
- *Biostatistics Practicum* — workflow infrastructure.
- *Statistical Computing in the Age of AI* — introductory methods and computing.
- *Advanced Statistical Computing in the Age of AI* — advanced numerical and Bayesian computation.
- *Applied Generative AI for Health Sciences Research* — generative AI as the orthogonal axis.
- *Applied Statistical Methods for Health Sciences Research* (this volume) — the applied-methods axis.

See the Preface for motivation and the Conventions page for visual cues.

License

Applied Statistical Methods for Health Sciences Research

A graduate textbook.

Copyright

To cite this book, please use:

> Thomas, R. G. (2026). *Applied Statistical Methods for Public Health*. Available at https://applied-methods.rgtlab.org.

Table of contents

Preface

This volume covers the applied methodological core of an MS-level biostatistics curriculum: the methods that every practising biostatistician encounters and that every MS programme requires, but that the four sister volumes either treat at foundation level or omit deliberately.

The book is the response to a curriculum-gap analysis across 24 graduate biostatistics programmes in the US and Europe (documented in `docs/curriculum-gap-analysis.md`). The analysis identified six topics that appear as core or required-elective material in 10 or more programmes but are absent from the existing four-volume sequence: longitudinal and correlated data at applied depth, survival analysis at applied depth, causal inference for observational data, clinical trial design and analysis, epidemiologic methods, and missing data at methodological depth. This volume covers all six, plus mediation, meta-analysis, and advanced categorical-data methods.

What this book covers

The 12 chapters are organised in five parts:

1. **Foundations.** Estimands and the estimand-estimator-estimate chain; epidemiologic measures and study design.
2. **Causal inference.** Foundations (potential outcomes, DAGs, exchangeability); estimation (propensity scores, IPW, g-methods, IV, RD, sensitivity); mediation.
3. **Correlated data and time-to-event.** Longitudinal and correlated data at applied depth; survival analysis (Cox, competing risks, RMST, recurrent events).
4. **Clinical trials.** Design (ICH E9 R1 estimands, adaptive, pragmatic); analysis and reporting.

5. **Specialised methods.** Missing data at depth; meta-analysis and evidence synthesis; advanced categorical (ordinal, multinomial, log-linear).

What this book does not cover

The book deliberately omits topics treated in the sister volumes or in dedicated specialty texts:

- The R / programming foundation (see *R for Biostatistics: A One-Week Boot Camp*).
- Reproducibility infrastructure (see *Biostatistics Practicum*).
- Linear models, GLM, mixed models, basic survival as model classes (see *Statistical Computing in the Age of AI*).
- Numerical stability, MCMC depth, HPC, high-dimensional methods (see *Advanced Statistical Computing in the Age of AI*).
- Generative AI as a workflow component (see *Applied Generative AI for Public Health and Biostatistics*).

Specialty topics that appear in fewer programmes but each warrant book-length treatment elsewhere: statistical genetics (see Foulkes; Laird and Lange), spatial statistics (Moraga; Banerjee/Carlin/Gelfand), time series for public health (Shumway and Stoffer), infectious-disease modelling (Vynnycky and White), health economics (Briggs/Sculpher/Claxton), categorical data theory (Agresti).

How this book relates to its siblings

Read the boot camp and the practicum first, then the introductory SCAI for the modelling foundations. Then read this volume; the methods here build on the GLM / mixed-model / survival foundations from SCAI but extend them in the directions an applied biostatistician actually needs. *Applied GenAI* is the orthogonal axis and can be read in parallel. *SCAI Advanced* is the companion deep-computing volume that picks up where SCAI leaves off.

Chapter template

Each content chapter follows the established sequence-wide structure: Learning objectives, Orientation, The statistician's contribution, content sections (with collapsible Check-your-understanding callouts), Worked example, Collaborating with an LLM, Principle in use, Exercises, Further reading. The template puts human judgement and verification at the centre of every chapter, rather than treating them as afterthoughts.

Acknowledgements

The chapter list reflects the curriculum-gap analysis across 24 programmes. The methodological literature underwriting each chapter is rich; canonical references appear in each chapter's Further reading.

Conventions

This page summarises the visual conventions used throughout the book.

Code

R code appears in syntax-highlighted blocks. Output is prefixed with `#>` to make the boundary between input and output explicit:

```
mean(c(1, 2, 3, 4, 5))
#> [1] 3
```

Inline code is in `monospace`. Function calls always include parentheses (`mean()` rather than `mean`) so that they are unambiguously functions. Package-qualified calls (`dplyr::filter`) appear when the function is not universally known, when there is name-collision risk, or when the chapter is teaching package usage.

Callouts

Three callout types appear:

> **Tip**
>
> A small practical recommendation.

Check your understanding: example

A short question testing comprehension of the just-read material. Click to expand the answer.

Warning

A pitfall the reader may otherwise hit.

Cross-references

Within this book, sections, figures, and tables are referenced by their Quarto label (`@sec-monte-carlo-human`, `@fig-mcmc-trace`, `@tbl-comparison`). These resolve to clickable links in HTML and proper figure/table numbers in PDF.

References to the companion volumes *Statistical Computing in the Age of AI* and *Biostatistics Practicum* use **prose pointers** rather than Quarto cross-references, because cross-references do not resolve across separate books. For example: 'see the Optimisation chapter of the companion *Statistical Computing in the Age of AI* volume'.

Mathematical notation

Conventional notation throughout. Vectors are bold lower-case ($\mathbf{x}$); matrices are bold upper-case ($\mathbf{X}$); scalars and parameters are non-bold. Estimators carry hats ($\widehat{\theta}$). Sample size is n; parameter dimension is p.

Chapter structure

Every content chapter follows the same template:

1. **Learning objectives.** What you will be able to do after reading.
2. **Orientation.** A short prose framing.

3. **The statistician's contribution.** What no tool can automate. The judgements at the centre of the chapter.
4. **Content sections** with **Check-your-understanding** callouts at natural pauses.
5. **Collaborating with an LLM on the chapter topic.** Prompt / Watch for / Verification triples for AI assistance.
6. **Exercises.** The work.
7. **Further reading.** Where to go next on the topic.

The pattern repeats deliberately. By the third chapter you know where to find each component.

How to use this book

This short orientation chapter explains the chapter template, the cross-references to sibling volumes, and the conventions used throughout the book. The template is the same as the SCAI-advanced and applied-genai sister volumes.

Chapter template

Every content chapter follows the same nine-section structure:

1. **Learning objectives.** A bulleted list of capabilities the reader should have after working through the chapter.
2. **Orientation.** Two to four paragraphs of prose framing: what the chapter is, why it matters, how it relates to adjacent chapters and to the sister volumes.
3. **The statistician's contribution.** A front-loaded section articulating the judgements at the centre of the chapter that no large language model can make on the reader's behalf.
4. **Content sections.** The chapter's substantive material, broken into sections with descriptive headings. Collapsible *Check your understanding* callouts appear at natural pauses.
5. **Worked example.** A concrete worked example, ideally biomedical, that exercises the chapter's tools end to end.
6. **Collaborating with an LLM on *topic*.** Three prompt patterns paired with what to watch for and how to verify, specific to the chapter's content.
7. **Principle in use.** Three habits that define defensible work in this area.
8. **Exercises.** Five exercises ranging from short conceptual checks to extended applied work.

9. **Further reading.** Canonical, modern applied, and software-documentation pointers.

The template is identical to the SCAI-advanced and applied-genai sister volumes; readers familiar with those volumes can read this one in the same rhythm.

Conventions

Visual cues used throughout the book are described on the Conventions page. Code blocks default to R; Python and Stan snippets are labelled. LLM prompts are shown in fenced blocks with the prompt text and verification commentary.

Companion repositories and audit trail

The book repository at https://github.com/rgt47/applied-methods contains:

- `docs/curriculum-gap-analysis.md` — the survey across 24 graduate programmes that identified the topical gaps this volume fills.
- `references.bib` — the working bibliography.

The audit trail is intentional: subsequent revisions of the volume will be informed by tracking how MS-programme curricula evolve, and the survey document is the benchmark.

Part I.

Foundations

1. Foundations of Applied Health-Research Statistics

1.1. Learning objectives

By the end of this chapter you should be able to:

- Articulate the estimand-estimator-estimate chain and apply it to a real research question.
- Distinguish a target population from a sampled population, and recognise the inferential consequences of the gap.
- Categorise a study by its design (RCT, cohort, case-control, cross-sectional) and identify the characteristic biases each design is vulnerable to.
- Recognise when a research question demands a causal, associational, or descriptive answer, and choose the estimand to match.

1.2. Orientation

Most applied biostatistical work fails not because the analyst chose the wrong test but because the question was under-specified before the test was chosen. The estimand- estimator-estimate framework, articulated formally in ICH E9 R1 and now standard across regulatory biostatistics, forces the question into focus before any software runs. This chapter establishes that framework and the study-design taxonomy that the rest of the book builds on.

The chapter is foundational in the strict sense: the remaining 11 chapters all assume the reader can answer 'what is the estimand here?' before opening R. Causal inference (Chs 3-4), longitudinal analysis (Ch 6), clinical trials

(Chs 8-9), missing data (Ch 10), and meta-analysis (Ch 11) each become well-defined only once the estimand is named.

The framing here inherits the public-health emphasis of the volume. The clinical-trial estimand framework, the epidemiologic study-design taxonomy, and the basic descriptive-vs-causal-vs-predictive question classification are equally important for designing a trial, planning a cohort study, or reading a paper.

1.3. The statistician's contribution

Three judgements at the centre of foundational work cannot be delegated.

(Judgement 1.) The question precedes the estimand. A research question stated as 'does treatment X affect outcome Y' is under-specified until the population, the intervention, the comparator, the timeframe, and the handling of intercurrent events are pinned down. The biostatistician's first contribution is to refuse the under-specified question and produce the specified version. 'Does sodium-glucose co-transporter inhibitor treatment, started within 30 days of MI in adults aged 40-80 with EF below 40%, reduce the 90-day risk of cardiovascular death compared to standard care, treating treatment discontinuation as part of the assigned strategy' is the question. The bare 'does X affect Y' is a draft.

(Judgement 2.) The estimand precedes the model. Two analyses of the same data with different estimands produce different numbers, and both are right answers to their own question. ITT vs. per-protocol analyses estimate different things; conditional vs. marginal effects estimate different things; the average treatment effect on the treated vs. the average treatment effect in the whole population estimate different things. The biostatistician picks the estimand by reasoning about what the report will inform, what action a reader could take after seeing the number, and writes the estimand into the analysis plan before fitting any model.

(Judgement 3.) The study design and the question must match. A randomised controlled trial answers a causal question (effect of assignment); a cohort study answers a question about association in the population followed (useful for surveillance, often a poor surrogate for the causal question); a case-control study answers a question about the odds of exposure

given outcome (rarely the question of substantive interest, often forced by feasibility). The biostatistician identifies the mismatch between the available design and the intended question and either redesigns or limits the claim accordingly.

These judgements distinguish work that informs decisions from work that produces plausible numbers in response to under-specified questions.

1.4. The estimand-estimator-estimate chain

The framework, as ICH E9 R1 (International Council for Harmonisation, 2019) formalises it for clinical trials and as the broader literature (Hernán & Robins, 2020; Lash et al., 2021) generalises for observational research:

Estimand. The thing you want to know. Five attributes pin it down:

1. **Population.** Who. The target population, defined by inclusion and exclusion criteria.
2. **Intervention.** What treatment, exposure, or condition is the subject of inference.
3. **Comparator.** What you are comparing the intervention against (often a reference treatment, placebo, or 'no intervention').
4. **Outcome.** What is measured, on whom, and when.
5. **Population-level summary.** How outcomes are summarised across the population: a difference of means, a risk ratio, a hazard ratio, an odds ratio.
6. **Intercurrent events.** What happens when patients discontinue, switch, or die before the outcome assessment, and how the analysis treats those events.

ICH E9 R1 specifies five strategies for handling intercurrent events:

- **Treatment policy.** Treat the intercurrent event as part of the assigned treatment strategy. ITT is a treatment-policy estimand.
- **Composite.** Treat the intercurrent event as part of the outcome (e.g., 'death or treatment failure').
- **Hypothetical.** Estimate what would have happened in a counterfactual world without the intercurrent event.

- **Principal stratum.** Estimate the effect in the subgroup defined by the (counterfactual) absence of the intercurrent event.
- **While-on-treatment.** Estimate the effect during the on-treatment period only.

Each strategy is a different estimand, with a different estimator and different assumptions. Two analyses of the same trial that use different strategies are answering different questions.

Estimator. The procedure for producing a number from data. A simple difference of means, a stratified difference of means, a Cox proportional hazards model, a propensity-score-weighted ATE estimator. Choosing the estimator is a separate decision from choosing the estimand; the same estimand can be estimated by several different estimators with different efficiency and robustness properties.

Estimate. The actual number you compute, with its uncertainty (standard error, confidence interval). The estimate is what appears in the report.

The chain reads: estimand to estimator to estimate. Reverse order is the failure pattern: the analyst computes a number (estimate), justifies it post-hoc as the answer to whatever question it nicely answers (estimand), and moves on. Estimand-first work prevents this.

Check your understanding: estimand vs. estimator

Question. Two statisticians analyse the same RCT. One uses ITT (everyone analysed in the assigned arm), the other uses per-protocol (only patients who actually received their assigned treatment). They produce different numbers. Which one is correct?
Answer.
Both can be correct, because they are estimating different estimands. ITT estimates the **treatment-policy** estimand: the effect of being assigned to a strategy, regardless of adherence. This is the answer to 'if I tell my patients to take this treatment, what happens?' Per-protocol estimates a quantity closer to the **hypothetical** estimand: what would happen if everyone adhered. This is the answer to 'if I could perfectly enforce the treatment, what happens?' Both questions are interesting; both have valid answers; the answers differ. The mistake is to report

one without naming which question it answers. The careful analyst reports both, identifies which is the primary estimand per the protocol, and discusses the discrepancy as informative about adherence.

1.5. Target population vs. sampled population

The **target population** is the population to which you want to generalise: for example, all adults aged 40-80 with type 2 diabetes in the United States. The **sampled population** is the population from which your data actually arose: for example, patients in three specific health systems in 2018-2024.

The gap matters. A model fit on the sampled population estimates the parameter in that population. Generalising to the target population requires:

- An argument that the sampled population resembles the target on the variables that matter for the question.
- Where the resemblance fails, an explicit generalisation step (post-stratification, transport formulas, or sensitivity analyses).
- Honest disclosure when generalisation is not defensible.

Biostatisticians frequently work on convenience samples (EHR data from one institution, a cohort that consented to research, the trial's centres). The temptation to generalise without justification is large; the discipline is to either justify or qualify.

1.6. Study-design taxonomy

Each design answers a different question and is vulnerable to different biases.

Randomised controlled trial (RCT). Subjects are randomised to intervention or comparator; outcomes are followed. Randomisation balances unmeasured confounders in expectation; the causal effect of assignment is identified. Vulnerable to: chance imbalance (especially with small samples), differential dropout, non-adherence, lack of blinding.

Cohort study. Subjects are followed forward in time; exposure is observed (not assigned); outcomes are recorded. Causal interpretation requires no unmeasured confounding, an assumption rarely defensible without auxiliary identification (instrumental variables, sensitivity analyses, design-based arguments). Suitable for surveillance, predictive modelling, and causal analysis with strong design or rich confounder data.

Case-control study. Subjects are sampled by outcome status; exposure is assessed retrospectively. Estimates the odds of exposure given outcome; estimable from data even when the outcome is rare. Causal interpretation is the hardest of the three designs and demands close attention to selection and recall biases.

Cross-sectional study. Subjects are surveyed at one time point. Estimates prevalence and associations; causal interpretation is impossible (the temporal order of exposure and outcome is unknown).

Quasi-experimental designs. Regression discontinuity, difference-in-differences, interrupted time series, synthetic control. Each exploits a feature of the intervention's roll-out (a sharp threshold, a phased introduction) to identify a causal effect under design- specific assumptions. The applied econometrics literature (Angrist & Pischke, 2009) is the natural home for this material.

The first question after 'what is the estimand?' is 'what design is the data from, and is the estimand identified by that design?' If the answer is no, either the estimand changes (to one identifiable from the design) or the analysis becomes a design-based argument plus sensitivity analyses.

1.7. Descriptive, associational, predictive, causal

A useful categorisation of question types (Hernán, 2018):

Descriptive. What is the prevalence of diabetes in this cohort? What proportion of patients have hbA1c $> 8\%$? Descriptive questions estimate population quantities; the inference is generalisation from the sampled to the target population.

Associational. Does hbA1c correlate with BMI in this cohort? Are diabetic patients more likely to be hypertensive? Associational questions estimate

joint or conditional distributions; no claim about why the association exists is made.

Predictive. Given a patient's age, BMI, and biomarker panel, what is the probability they will have a CV event in 5 years? Predictive questions optimise out-of-sample accuracy; the model need not be causal.

Causal. If we assigned this patient to treatment X rather than treatment Y, would their CV-event probability change? Causal questions estimate counterfactual contrasts; the inference requires assumptions about confounding, exchangeability, and positivity.

Different questions need different methods. A model that is excellent for prediction may be a poor causal estimator; a model that gives valid causal estimates may predict poorly. The biostatistician identifies the question category early and chooses methods to match.

1.8. Worked example: specifying an estimand for an observational study

A clinical team has access to electronic health records from three hospitals (2018-2024) and wants to know 'whether SGLT2 inhibitors are effective in real-world patients with heart failure'. They ask the biostatistician to plan the analysis.

The biostatistician's first move is to refuse the question as stated and write the specified version.

Population. Adults aged 18+ with a confirmed diagnosis of heart failure with reduced ejection fraction (HFrEF, EF $< 40\%$) seen at any of the three hospitals in 2018-2023. Exclude patients with end-stage renal disease, type 1 diabetes, or pregnancy.

Intervention. Initiation of an SGLT2 inhibitor within 30 days of HFrEF diagnosis.

Comparator. No SGLT2 inhibitor initiation within 30 days of diagnosis.

Outcome. All-cause mortality at 12 months from diagnosis.

Population-level summary. Risk ratio of mortality (intervention vs. comparator), with 95% CI.

Intercurrent events. Initiation of SGLT2 in the comparator group after 30 days: treatment-policy strategy (analyse as comparator). Treatment discontinuation in the intervention group: treatment-policy strategy. Death from causes other than HF: contributes to the outcome.

Design. Cohort study, observational. Causal interpretation requires no unmeasured confounding given the available baseline covariates (age, sex, EF, NT-proBNP, eGFR, comorbidities, baseline meds). Confounder-adjusted using inverse-probability weighting (Ch 4); sensitivity analysis via E-value (Ch 4).

Generalisation. Estimates apply to the three-hospital sampled population. Generalisation to the broader US HFrEF population requires an additional argument (comparison of the sampled population's age, sex, race, comorbidity distribution to a national reference) provided in the discussion.

The biostatistician's two-page protocol locks in the estimand before any data is queried. This protocol is what the analysis plan, the report, and the published paper will all defer to. Six months later, when a collaborator asks 'why did you use 30 days as the window?' or 'why is the comparator group not just non-users?', the protocol is the answer.

1.9. Collaborating with an LLM on applied-methods foundations

Three patterns for using AI assistance well at the foundation stage.

Prompt 1: 'Write the estimand for this research question.' Provide a one-paragraph informal description of the question and the available data.

What to watch for. The LLM produces a competent estimand statement that captures the obvious five attributes. It commonly under-specifies the intercurrent-event handling and the temporal definitions ('within how many days?'). Ask follow-up questions about each attribute it answered tersely.

Verification. Read the LLM's estimand against ICH E9 R1 if you have it; check that all five attributes are specified concretely (not just named). For each intercurrent event, the LLM should pick a strategy explicitly.

Prompt 2: 'What study designs would identify this estimand?' Provide the estimand and the constraints (retrospective vs. prospective, available data, ethical constraints).

What to watch for. The LLM generally identifies the correct designs but tends to undersell the assumptions required for observational identification. Push back: 'what specific unmeasured confounders would invalidate this analysis?' is a useful follow-up.

Verification. The LLM-suggested design is a candidate; your own knowledge of the data and the literature is the veto. The LLM has not seen your data and does not know which confounders are measured.

Prompt 3: 'Translate this question into a descriptive/associational/predictive/causal classification.' Provide the question.

What to watch for. The LLM is reasonably good at this but tends to hedge ('this could be associational or causal depending on…'). Push for a single answer plus a brief justification.

Verification. Compare the LLM's classification to your own. Disagreement is informative: it usually indicates the question itself is ambiguous and worth revising.

The meta-pattern: LLMs are useful for generating drafts of estimands and study-design proposals, and for forcing you to articulate things you might have left implicit. They are not substitutes for the biostatistician's domain judgement. The right use is 'AI drafts, statistician edits and approves'.

1.10. Principle in use

Three habits define defensible foundational work.

1. **Write the estimand before fitting the model.** The estimand belongs in the analysis plan, not in the results section. The analysis plan is the contract; the model is the procedure that satisfies it.

2. **Match the estimand to the design.** If the data come from a cohort, do not write an estimand that only an RCT can identify. The mismatch is the biostatistician's responsibility to surface.
3. **Disclose the gap between sampled and target populations.** Every report should contain a one-paragraph statement about who the estimates apply to and why.

1.11. Exercises

1. For a research question of your choice (a study you are working on, or a recent paper in your field), write the estimand using all six ICH E9 R1 attributes. Identify the intercurrent-event strategies for at least two intercurrent events.

2. Take a paper from a recent issue of *NEJM* or *Lancet*. Identify the estimand the paper claims to have estimated. Identify the actual analysis. Are they the same? Where do they diverge?

3. For a published cohort study in your area, list the confounders the authors adjusted for. Now list the confounders they should have adjusted for given biological plausibility. What is the gap, and what is the likely direction of bias?

4. Draft the descriptive-vs-associational-vs-causal classification for ten research questions from your field. Where the classification is ambiguous, write one sentence specifying what would resolve the ambiguity.

5. For one estimand from problem 1, identify three different estimators that could compute it. List the assumptions each requires and the typical efficiency ranking.

1.12. Further reading

- International Council for Harmonisation (2019), *ICH E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials*. The regulatory document that anchors the estimand framework.

- Hernán & Robins (2020), *Causal Inference: What If*. The open-access textbook that develops the estimand-and- identifiability discipline for observational data.
- Lash et al. (2021), *Modern Epidemiology* (4th edition). The reference textbook for the epidemiologic study- design taxonomy.
- Hernán (2018), 'The C-word: Scientific euphemisms do not improve causal inference from observational data'. The argument for naming causal questions as causal.

2. Epidemiologic Measures and Study Design

2.1. Learning objectives

By the end of this chapter you should be able to:

- Compute and interpret the standard epidemiologic measures: prevalence, incidence rate, cumulative incidence, risk ratio, odds ratio, rate ratio.
- Choose between cohort, case-control, and cross-sectional designs for a given research question, and articulate the tradeoffs.
- Distinguish confounding, effect modification, selection bias, and information bias, and recognise each in applied work.
- Apply the modern target-trial-emulation framework to a proposed observational analysis.

2.2. Orientation

Chapter 1 established estimands and study designs at the framework level. This chapter operationalises them: how to compute the standard measures, how to choose designs, how to recognise the biases that threaten each design's validity. The material is the entry point to applied epidemiology and the conceptual scaffolding for the causal-inference chapters that follow.

The chapter is organised around three threads. **Measures**: prevalence, incidence, and the rate / risk / odds ratios that summarise associations and effects. **Designs**: cohort, case-control, cross-sectional, with explicit attention to when each is the right choice. **Biases**: confounding, selection, information, with examples drawn from public-health practice.

Modern epidemiology has converged on **target-trial emulation** (Hernán & Robins, 2016) as the discipline for observational causal inference. The target-trial framework asks the analyst to specify the hypothetical randomised trial that would answer the question, and then to design the observational analysis to emulate that trial as closely as possible. The chapter ends by walking through this framework on a concrete example.

2.3. The statistician's contribution

Three judgements at the centre of epidemiologic design are not delegable.

(Judgement 1.) Choose the measure that matches the decision. A risk ratio of 1.5 sounds different from a risk difference of 0.05, but they may describe the same data. The right measure is the one that informs the decision under consideration: relative measures for mechanism and aetiology questions; absolute measures (risk differences, NNT) for clinical and policy decisions where the baseline risk matters. The biostatistician chooses the measure deliberately and presents the alternative when the decision benefits.

(Judgement 2.) The design constrains the inference. A case-control study cannot estimate prevalence; a cross-sectional study cannot establish temporal order; an unmatched cohort is vulnerable to confounding by indication. The biostatistician identifies what the chosen design can and cannot say, and limits the report's claims accordingly. Reviewers will identify the limit if the analyst does not; better to disclose than to defend.

(Judgement 3.) Bias is not a hypothesis test. The question is not 'is there confounding' (the answer is almost always yes) but 'how large is the bias relative to the effect, and in which direction'. The biostatistician quantifies bias through sensitivity analyses (E-values, tipping-point analyses, negative controls) rather than asserting its absence. The discipline applies as much to RCTs (where confounding is in expectation zero but in any specific trial may be present) as to observational studies.

These judgements distinguish epidemiologic work that informs decisions from work that produces numbers in search of a question.

2.4. Prevalence and incidence

Prevalence is the proportion of a population with the condition at a point in time:

$$P = \frac{\text{cases at time } t}{\text{population at time } t}$$

Prevalence is dimensionless. It depends on incidence (how often the condition is acquired) and duration (how long it lasts). High-prevalence conditions can have either high incidence or long duration or both.

Incidence rate is the number of new cases per person-time:

$$\text{IR} = \frac{\text{new cases in period}}{\text{person-time at risk}}$$

Person-time has units (person-years, person-months). Incidence rate has the corresponding inverse units (per person-year). It is the rate parameter of the underlying point process.

Cumulative incidence is the proportion of an initially-disease-free cohort that develops the condition over a defined period:

$$\text{CI}(t) = \Pr(T \leq t)$$

Cumulative incidence is dimensionless. For short follow-up with a constant rate, $\text{CI}(t) \approx \text{IR} \cdot t$; the approximation breaks at longer follow-up. The full relationship via the survival function (Ch 7) is $\text{CI}(t) = 1 - S(t)$.

The three measures answer different questions: prevalence tells you the burden at a moment; incidence rate tells you how fast new cases appear; cumulative incidence tells you the total proportion affected over a follow-up period. Reports often confuse these; distinguish them carefully.

2.5. Risk, rate, and odds ratios

Three ratios summarise associations between an exposure and an outcome.

Risk ratio (RR): ratio of cumulative incidence between exposed and unexposed.

$$\text{RR} = \frac{\text{CI}_{\text{exposed}}}{\text{CI}_{\text{unexposed}}}$$

Rate ratio (or hazard ratio in survival contexts): ratio of incidence rates.

$$\text{IRR} = \frac{\text{IR}_{\text{exposed}}}{\text{IR}_{\text{unexposed}}}$$

Odds ratio (OR): ratio of odds of disease.

$$\text{OR} = \frac{p_{\text{exposed}}/(1-p_{\text{exposed}})}{p_{\text{unexposed}}/(1-p_{\text{unexposed}})}$$

When the outcome is rare ($p \ll 0.1$), OR $\approx$ RR. For common outcomes, OR exaggerates effect size relative to RR; reporting only OR for a common outcome can mislead.

The choice between measures depends on the design and the question:

- **RCT or cohort, common outcome**: prefer RR or risk difference. Logistic regression's OR is a software default but not always the right reporting choice.
- **Case-control**: OR is what the design naturally estimates; under the rare-disease assumption it approximates the RR.
- **Cohort with time-to-event**: rate ratio or hazard ratio (Ch 7).

Check your understanding: OR vs. RR for common outcomes

Question. A cohort study estimates an OR of 2.5 for the association between a binary exposure and a binary outcome. The unexposed risk is 0.30. What is the RR?
Answer.
The unexposed odds is $0.30/0.70 = 0.43$; the exposed odds is $2.5 \times 0.43 = 1.07$; the exposed risk is $1.07/(1+1.07) = 0.52$. The RR is $0.52/0.30 = 1.73$, much smaller than the OR of 2.5. For a common outcome, reporting the OR makes the effect look larger than it is. The RR is the more interpretable measure here.

2.6. Cohort studies

Subjects are sampled by exposure (or unselectively) and followed forward in time; outcomes are recorded as they occur. Strengths:

- Establishes temporal order (exposure precedes outcome).
- Estimates absolute risks and rates directly.
- Suitable for causal inference under a no-unmeasured-confounding argument plus appropriate adjustment.

Weaknesses:

- Expensive and slow; long follow-up to accrue events for rare outcomes.
- Loss to follow-up creates differential bias if loss correlates with exposure or outcome.
- Confounding by indication when exposure is treatment-like.

Cohorts can be **prospective** (assembled at exposure, followed forward) or **retrospective** (assembled from records that already exist, with exposure and outcome already observed). Retrospective cohorts use the prospective machinery on records data; the analyst must take care that exposure was recorded before outcome and not produced by knowledge of the outcome.

2.7. Case-control studies

Subjects are sampled by outcome status: cases (with the disease) and controls (without). Exposure is then assessed retrospectively. Strengths:

- Efficient for rare outcomes (cases are oversampled relative to their population frequency).
- Multiple exposures can be studied for the same outcome.
- Often the only feasible design for slow-developing conditions.

Weaknesses:

- Recall bias when exposure is self-reported (cases may remember exposure differently from controls).
- Selection bias is a constant threat: how were the controls chosen?

- Estimates only the OR; no absolute risks.

The OR from a case-control study approximates the RR in the underlying population only when the outcome is rare. For common outcomes, the OR is the parameter the design naturally estimates and should be reported as such.

Nested case-control designs sample cases and a matched set of controls from within an existing cohort, combining the efficiency of case-control with the defensibility of cohort sampling.

2.8. Cross-sectional studies

A snapshot of a population at one time point. Estimates prevalence and prevalence ratios. Cannot establish temporal order: if exposure and outcome are both observed at one time, you cannot tell which preceded.

Useful for surveillance and prevalence estimation; less useful for inference about causes (the temporal-order problem is fundamental, not addressable by adjustment). Many published cross-sectional studies overstate causal claims; the careful reader treats their associations as hypothesis-generating, not hypothesis-testing.

2.9. Confounding, effect modification, selection, information

Four threats to validity. Each requires a different response.

Confounding is a third variable that affects both exposure and outcome and biases the exposure-outcome association. The classic test: smoking confounds the coffee-lung-cancer association because smokers drink more coffee and smokers get lung cancer at higher rates. Adjustment, restriction, matching, or design (randomisation) addresses confounding. The identification assumption is no unmeasured confounding given the adjustment set; sensitivity analysis quantifies what unmeasured confounding would need to do to overturn the conclusion.

Effect modification (interaction) is when the effect of the exposure differs across levels of a third variable. Sex modifies the effect of cardiovascular medications; age modifies the effect of vaccinations. Effect modification is a feature of the data, not a bias; it is reported by stratification or interaction terms in the model.

Selection bias arises when inclusion in the study depends on both exposure and outcome. The classic example: hospital-based controls in a case-control study, where hospitalisation is itself associated with the exposure. Selection bias cannot be fixed by adjustment; it is addressed by design (sampling controls appropriately) or by sensitivity analysis (quantifying the magnitude of bias under plausible selection mechanisms).

Information (measurement) bias arises when exposure or outcome is mismeasured, and the mismeasurement correlates with the other variable. Differential misclassification is the worst case: cases recall exposure better than controls. Non-differential misclassification (random measurement error) generally biases toward the null. Validation studies and quantitative-bias analyses address information bias.

The four are distinct. A common error is to treat 'confounding' as a catch-all for 'something is wrong with this association'; precision in naming the bias informs the right response.

2.10. Target-trial emulation

The discipline that has emerged for observational causal inference (Hernán & Robins, 2016, 2020):

1. **Specify the target trial.** What randomised trial would answer the question if it could be run? Specify eligibility, treatment strategies, assignment, follow-up, outcome, intercurrent events.
2. **Identify the observational data that emulates the trial.** Eligibility maps to inclusion at time zero; treatment strategies map to observed exposure patterns; assignment maps to the analyst's adjustment for confounding.
3. **Document the emulation.** Where the observational data fails to emulate the target trial (e.g., immortal-time bias from how exposure

is defined in the records), name the failure and address it (typically by redefining time zero so exposure and eligibility are simultaneous).

The framework eliminates a class of common observational-design errors: immortal-time bias, prevalent-user bias, selection-on-treatment-after- baseline. Most published observational studies of drug effects can be substantially improved by applying the target-trial framework retroactively. The discipline has become standard practice in modern pharmacoepidemiology.

2.11. Worked example: target-trial emulation for SGLT2 inhibitors

The example from Chapter 1 (SGLT2 effectiveness in HFrEF, EHR data from three hospitals) is a textbook target-trial application.

Step 1. Specify the target trial. A hypothetical RCT enrols adult HFrEF patients within 30 days of diagnosis, randomises to SGLT2 vs. no SGLT2 (treatment policy: any subsequent treatment changes are followed under the assigned arm), follows for 12 months, and records all-cause mortality.

Step 2. Emulate. From the EHR cohort:

- **Eligibility.** Adults aged 18+, HFrEF (EF $< 40\%$) diagnosed during the study period. Exclude prior SGLT2 users (who would not be eligible for the hypothetical trial).
- **Time zero.** Date of HFrEF diagnosis.
- **Treatment groups.** 'SGLT2 initiator' = SGLT2 prescription within 30 days of time zero. 'Non-initiator' = no SGLT2 prescription within 30 days. Note: this requires defining the groups at time zero, before observing the treatment decision; common in target-trial emulation, awkward without it.
- **Follow-up.** From time zero to 12 months, death, or end of records.
- **Confounding.** Adjust for baseline confounders at time zero (age, sex, EF, NT-proBNP, eGFR, comorbidities, baseline medications). Use IPW (Ch 4).
- **Sensitivity.** E-value for unmeasured confounding (Ch 4).

Step 3. Identify failures of emulation.

- Patients who initiate SGLT2 between days 30 and 90 are misclassified as non-initiators in the observational data but would be in the initiator arm under treatment policy in the target trial. The target-trial framework calls this 'misclassification of grace period'; the fix is either to extend the grace period or to use a clone-censor-weight approach.
- Time zero is hard to define for patients diagnosed at outside hospitals and transferred in. Address by either restricting to in-network diagnoses or using a sensitivity analysis.

The target-trial protocol is now the analysis plan. Six months later, a reviewer asking 'why this 30-day window?' is answered by 'the target trial defined treatment groups at time zero with a 30-day grace period; the alternative would be a different target trial'.

2.12. Collaborating with an LLM on epidemiologic study design

Three patterns that work.

Prompt 1: 'Write the target trial for this observational study.' Provide the research question and the data source.

What to watch for. The LLM produces a competent draft of the seven target-trial components (eligibility, treatment, assignment, follow-up, outcome, intercurrent events, analysis). It tends to under-specify the treatment-strategy definition and the time-zero question. Push back: 'how would you define time zero given that exposure is observed in records, not assigned at time zero?'

Verification. Read the LLM's target trial against Hernán & Robins (2016); check that all seven components are specified concretely and that the immortal-time issue is addressed.

Prompt 2: 'Identify the threats to validity in this analysis.' Provide the analysis description.

What to watch for. The LLM lists the standard biases (confounding, selection, measurement) but tends to be generic. Push for specific confounders, specific selection mechanisms, specific measurement-error patterns informed by the data source.

Verification. The LLM's threats are a starting list; your knowledge of the data source is what specialises them. Add threats the LLM missed; remove ones that do not apply.

Prompt 3: 'Compute the relevant epidemiologic measures from this 2x2 table.' Provide the cell counts.

What to watch for. The LLM gets the arithmetic right for risk ratio, odds ratio, etc. It sometimes confuses the OR-RR distinction for common outcomes. Verify the specific case the LLM is computing.

Verification. Recompute by hand; the formulas are elementary.

The meta-pattern: LLMs are useful for drafting target-trial protocols and for listing the threats to validity in standard form. They cannot evaluate whether your specific data emulates the target trial well; that judgement is yours.

2.13. Principle in use

Three habits define defensible epidemiologic work.

1. **Match the measure to the decision.** Relative measures for aetiology; absolute measures for policy and clinical decisions. Report both when the audience benefits.
2. **Apply the target-trial framework.** Every observational causal analysis specifies its target trial and documents where the emulation succeeds or fails.
3. **Quantify bias rather than assert its absence.** Sensitivity analyses (E-values, tipping points) are part of every causal analysis, not an afterthought.

2.14. Exercises

1. For a published cohort study in your field, identify the prevalence, incidence rate, and cumulative incidence reported (or compute them from the reported data). Note the units of each.
2. Take a 2x2 contingency table (exposed-vs-unexposed by outcome-vs-no-outcome) of your choice. Compute the risk ratio, the odds ratio, the rate ratio (assuming person-time is proportional to sample size), and the risk difference. Discuss when each measure is the right choice for reporting.
3. Identify a published case-control study in your field. Identify how controls were sampled and whether the sampling could introduce selection bias. If yes, describe the likely direction.
4. For an observational study of a drug effect, write the target-trial protocol following the seven components. Identify two specific places where the observational data will fail to emulate the target trial perfectly, and propose a remediation for each.
5. Compute an E-value for an observed risk ratio of 1.6. What does the E-value tell you about the strength of unmeasured confounding required to explain away the association?

2.15. Further reading

- Lash et al. (2021), *Modern Epidemiology* (4th edition). The reference textbook for the material in this chapter.
- Hernán & Robins (2020), *Causal Inference: What If*. The open-access textbook that develops target-trial emulation in depth.
- Hernán & Robins (2016), 'Using big data to emulate a target trial when a randomized trial is not available'. The methods paper that defines the target-trial framework.
- VanderWeele & Ding (2017), 'Sensitivity analysis in observational research: introducing the E-value'. The reference for E-value sensitivity analysis.

Part II.

Causal Inference

3. Causal Inference I: Foundations

3.1. Learning objectives

By the end of this chapter you should be able to:

- Articulate the potential-outcomes framework and use counterfactual notation to define the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the conditional average treatment effect (CATE).
- Draw a directed acyclic graph (DAG) for a research question and identify the minimal sufficient adjustment set using the backdoor criterion.
- State the three core identifying assumptions (exchangeability, positivity, consistency) and recognise where each is violated in applied work.
- Distinguish association from causation in language and in symbols, and recognise common conflations.

3.2. Orientation

Causal inference is the discipline of making counterfactual claims from observed data: 'if this patient had been assigned the treatment, their outcome would have been different.' The discipline has a precise mathematical foundation (Rubin's potential-outcomes framework, Pearl's structural causal models) and a small set of identifying assumptions that determine when counterfactual claims are estimable from data.

This chapter establishes the foundations: the potential-outcomes notation, the average treatment effects of interest, the DAG vocabulary, the backdoor

criterion, and the three core assumptions. Chapter 4 turns the foundations into estimation procedures (propensity scores, IPW, g-methods).

The chapter is shorter than its companion (Ch 4) because the conceptual investment is the binding constraint. A reader who internalises the potential-outcomes notation and the backdoor criterion can read modern causal-inference literature; a reader who has not will struggle with the next chapter.

3.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Distinguish causal from associational language. The English-language word 'effect' is ambiguous; the technical word is not. A statement like 'smoking is associated with lung cancer' is associational; 'smoking causes lung cancer' is causal. The first is a fact about a population's joint distribution; the second is a counterfactual claim that requires identifying assumptions. The biostatistician keeps the language precise: associational claims for descriptions, causal claims only when the assumptions are stated and defensible.

(Judgement 2.) The DAG is a domain claim. A directed acyclic graph encoding the causal relationships in your problem is a substantive scientific claim, not a statistical artefact. Drawing the DAG forces you to articulate which variables affect which others; the arrows are the claim. The biostatistician engages domain experts to verify the DAG. A DAG with the arrows wrong produces an analysis that is precisely incorrect; the precision is no help.

(Judgement 3.) Identifying assumptions are substantive, not statistical. The no-unmeasured-confounding assumption (exchangeability) is not testable from data alone; it is a claim about what is in the world, defended by the design and the scientific argument. The biostatistician's role is to make the assumption explicit, defend it where defensible, and quantify the consequences of violation where it is not.

These judgements are what distinguish a causal analysis from a regression model that happens to use causal vocabulary.

3.4. The potential-outcomes framework

Notation (Neyman, 1923; Rubin, 1974):

- A: the treatment indicator ($0 =$ no treatment, $1 =$ treated; or generally, the value taken by an intervention).
- Y: the observed outcome.
- $Y(a)$: the potential outcome under treatment level a. A patient has potential outcomes $Y(0)$ and $Y(1)$; the observed outcome is $Y = Y(A)$ — whichever potential outcome corresponds to the treatment actually received.
- X: a vector of pre-treatment covariates.

The fundamental problem of causal inference is that each individual has potential outcomes $Y(0)$ and $Y(1)$, but only one is observed for any given person. The other is the **counterfactual** outcome — what would have happened if the treatment assignment had been different.

Causal effects are contrasts of potential outcomes. The **individual treatment effect**:

$$\tau_i = Y_i(1) - Y_i(0)$$

is fundamentally unobservable for any one individual. The **average treatment effect** (ATE):

$$\text{ATE} = E[Y(1) - Y(0)]$$

is identifiable from observed data under appropriate assumptions.

Two related estimands matter:

The **average treatment effect on the treated** (ATT):

$$\text{ATT} = E[Y(1) - Y(0) \mid A = 1]$$

the average effect among those actually treated.

The **conditional average treatment effect** (CATE):

$$\text{CATE}(x) = E[Y(1) - Y(0) \mid X = x]$$

the average effect at a specific covariate value.

ATE, ATT, and CATE are different things and often have different values. Pick the one that matches the decision: ATE for a population-level policy ('should we offer the treatment universally'), ATT for a question about the actually-treated population ('among those who got the treatment, what was the average effect'), CATE for personalised medicine ('what is the effect for this kind of patient').

Check your understanding: ATE vs. ATT

Question. A new diabetes drug is mostly prescribed to patients with poor glycaemic control on existing therapy. The ATT (effect among those treated) is large and positive. The ATE (average effect across all diabetic patients) is smaller. Which is the relevant quantity for the decision 'should the drug be added to guideline-recommended treatment for all diabetic patients'?

Answer.

The ATE is the relevant quantity. The ATT applies only to the historical pattern of treatment (poorly-controlled patients); guideline change would treat patients across the whole spectrum, including better-controlled ones who were not represented in the ATT. The ATE captures the expected average effect under universal treatment. Reporting only the ATT for a guideline-change question overstates the expected benefit.

3.5. The three core assumptions

Three identifying assumptions allow ATE to be computed from observed data.

(A1) Exchangeability (no unmeasured confounding):

$$Y(a) \perp\!\!\!\perp A \mid X \quad \text{for all } a.$$

Conditional on X, the potential outcomes are independent of treatment assignment. Equivalently: treatment assignment is as good as random within strata of X.

This is the load-bearing assumption. In an RCT, randomisation makes it true by design. In observational studies, it is a substantive claim about the

world; it fails when there is a confounder U that affects both A and Y that is not in X.

(A2) Positivity:

$$0 < \Pr(A = a \mid X) < 1 \quad \text{for all } a, x.$$

Every covariate combination has some probability of receiving each treatment level. If certain combinations are deterministically untreated (or deterministically treated), the data carry no information about the counterfactual at that combination.

Positivity violation is real: in observational drug data, certain patient subgroups are essentially never prescribed certain drugs (contraindications, age cutoffs, specialist referral patterns). The analyst must check the propensity-score distribution and either restrict the analysis or impose strong modelling assumptions to extrapolate.

(A3) Consistency:

$$A = a \implies Y = Y(a).$$

The observed outcome under a particular treatment is the same as the potential outcome under that treatment. This sounds tautological but encodes a substantive claim: there is a single, well-defined version of each treatment, and the treatment received is the treatment of interest.

Consistency fails when the treatment is heterogeneous ('SGLT2 inhibitors' might mean dapagliflozin or empagliflozin, with different effects) or when the analyst defines treatment in a way that does not match the observed exposure. The fix is precision in the treatment definition.

Identification under A1-A3: When all three hold, the ATE is identified from data:

$$\text{ATE} = E[E[Y \mid A = 1, X] - E[Y \mid A = 0, X]].$$

The outer expectation is over the distribution of X; the inner is the difference in conditional outcome means between treatment groups.

This identification formula is the basis for g-computation (Ch 4); other estimators are equivalent under A1-A3 but use different computational routes.

3.6. Directed acyclic graphs

A DAG is a graphical representation of causal relationships:

- **Nodes** are variables.
- **Directed edges** ($A \to B$) mean A is a direct cause of B.
- **Acyclic** means no variable causes itself through any path.

DAGs are useful because they make causal assumptions explicit and allow algorithmic reasoning about identification.

A simple example: smoking (S) affects both coffee consumption (C) and lung cancer (Y). Coffee does not cause lung cancer. The DAG:

The association between C and Y is non-zero (they share a common cause) even though C does not cause Y. Adjusting for S removes the spurious association.

Three structural roles a variable can play on a path between A and Y:

Confounder ($A \leftarrow Z \to Y$): Z is a common cause of both. Adjustment for Z removes the non-causal association.

Mediator ($A \to M \to Y$): M is on the causal path from A to Y. Adjustment for M blocks part of the causal effect; this is desired in mediation analysis (Ch 5) but not in the basic ATE.

Collider ($A \to C \leftarrow Y$): C is a common effect. Adjustment for a collider creates spurious association (selection bias). Do not adjust for colliders.

The **backdoor criterion** (Pearl, 1995): X is a sufficient adjustment set for the causal effect of A on Y if (a) X blocks all backdoor paths from A to Y and (b) X contains no descendants of A.

A backdoor path is any path from A to Y that begins with an arrow into A. Confounders sit on backdoor paths and need to be adjusted for; mediators do not sit on backdoor paths and must not be adjusted for.

The DAG and the backdoor criterion together produce a recipe for identification: draw the DAG, identify the backdoor paths, find an adjustment set that blocks all of them. The set is the variables to put in the model.

The `dagitty` R package implements the algorithm; for a given DAG and exposure-outcome pair it returns the minimal sufficient adjustment sets:

```
library(dagitty)
g <- dagitty('dag {
  smoking -> coffee
  smoking -> lung_cancer
  coffee -> lung_cancer [adjusted = "no"]
}')
adjustmentSets(g, exposure = "coffee",
               outcome = "lung_cancer")
#> { smoking }
```

The output: to identify the (zero) effect of coffee on lung cancer, adjust for smoking.

3.7. When the assumptions fail

Each assumption fails in characteristic ways and demands a characteristic response.

Exchangeability fails (unmeasured confounding). The standard responses:

1. **Sensitivity analysis.** Quantify how strong an unmeasured confounder would have to be to overturn the conclusion. The E-value (VanderWeele & Ding, 2017) is the standard summary; it answers 'how strong does the unmeasured confounding need to be to explain the observed association.'

2. **Negative controls.** A negative-control exposure (one expected to have no effect on the outcome) and a negative-control outcome (one expected not to be affected by the exposure) check for residual confounding. Non-zero estimates on negative controls suggest residual confounding that affects the primary analysis too.
3. **Instrumental variables.** Find a variable that affects exposure but only affects the outcome through exposure. Identifies a different estimand (the effect among 'compliers') under different assumptions; covered in Ch 4.
4. **Design-based identification.** Regression discontinuity, difference-in-differences, synthetic control. Each exploits a feature of the intervention's roll-out for identification.

Positivity fails. The standard responses:

1. **Restrict the population** to the region of covariate support that has positive probability of each treatment.
2. **Trim the propensity score** to remove observations near 0 or 1.
3. **Use a doubly robust estimator** that is more robust to extreme weights (Ch 4).

Consistency fails (treatment heterogeneity). The standard response: redefine the treatment more precisely. 'SGLT2 use' is too coarse; 'dapagliflozin 10mg/day for at least 30 days starting within 30 days of HFrEF diagnosis' is the kind of precision consistency requires.

3.8. Worked example: drawing a DAG

A research question: does parental income affect adolescent academic performance, controlling for the right confounders?

Available variables: parental income, adolescent academic performance, parental education, neighbourhood socio-economic status, school quality, parental involvement, adolescent IQ.

Drawing the DAG:

```
parental_education
      |
      v
parental_income ---> adolescent_performance
      ^                       ^
      |                       |
neighbourhood_ses --------> school_quality
                              ^
                              |
                        parental_involvement (mediator)
```

(In an actual paper, draw the graph; here the text suggests it.)

Backdoor analysis: backdoor paths from parental income to adolescent performance go through parental education and neighbourhood SES. The minimal adjustment set is `{parental_education, neighbourhood_ses}`.

Variables NOT to adjust for:

- Parental involvement (mediator on the income → involvement → performance path; adjusting blocks part of the causal effect of income).
- Adolescent IQ (downstream of the exposure if we believe income affects cognitive development).

The DAG forces these distinctions. Without it, an analyst might 'adjust for everything' (including mediators) and report a biased estimate.

3.9. Collaborating with an LLM on causal-inference foundations

Three patterns.

Prompt 1: 'Draw the DAG for this research question.' Provide the variables and the substantive claims about which causes which.

What to watch for. The LLM produces a plausible DAG but defaults to the most common pattern in similar research; it may miss domain-specific edges. Push for explicit confirmation of each edge ('does X cause Y? why?') and engage your domain expert.

Verification. Show the DAG to a domain expert. Edges that the LLM proposed but the expert rejects are substantive claims to remove. Edges the expert adds are substantive claims to add.

Prompt 2: 'Find the minimal adjustment set for this DAG.' Provide the DAG (in dagitty syntax or a description).

What to watch for. The LLM can apply the backdoor criterion correctly on simple DAGs and gets confused on DAGs with many nodes. Verify by running `dagitty` on the same DAG.

Verification. `dagitty::adjustmentSets()` is the ground truth.

Prompt 3: 'List the assumptions required for this analysis.' Provide the analysis description.

What to watch for. The LLM lists exchangeability, positivity, consistency, but tends to be generic about which specific confounders threaten exchangeability. Push for the specific case.

Verification. The LLM's list is a starting point; your knowledge of the data informs which assumptions are most at risk.

The meta-pattern: LLMs are good for the syntactic mechanics of causal inference (drawing DAGs, listing assumptions) and bad at the substantive ones (which edges go in the DAG, which assumptions are likely to hold). Use them for drafts; the substance is yours.

3.10. Principle in use

Three habits define defensible causal-inference work.

1. **Draw the DAG before fitting the model.** The DAG is the causal claim; the model is the procedure. A model fitted without a DAG is not a causal analysis, regardless of what the report says.
2. **State the three assumptions explicitly.** Every causal report names exchangeability, positivity, and consistency, and discusses how each is defended.
3. **Run a sensitivity analysis.** Every causal estimate has an E-value or equivalent; the discussion includes what unmeasured confounding would do.

3.11. Exercises

1. For a research question of your choice, write the ATE, ATT, and CATE in potential-outcomes notation. Identify which is the most decision-relevant for the intended audience.
2. Draw the DAG for a research question in your area. Identify the back-door paths and the minimal sufficient adjustment set. Use `dagitty` to check your answer.
3. For one of the three core assumptions, identify a specific way it could fail in a hypothetical observational study and propose a sensitivity analysis to address it.
4. Compute the E-value for a published RR of 2.0. What does this E-value tell you about the strength of unmeasured confounding required to overturn the conclusion?
5. Write a one-paragraph protocol section that states the estimand (using potential-outcomes notation), the identifying assumptions, the design that identifies the estimand, and the sensitivity analyses.

3.12. Further reading

- Hernán & Robins (2020), *Causal Inference: What If*. The reference textbook for the framework introduced in this chapter.
- Pearl (2009), *Causality: Models, Reasoning, and Inference*. The reference for the structural causal model and graphical reasoning.
- Rubin (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies'. The foundational paper for potential outcomes.
- The `dagitty` R package and online tool (https://www.dagitty.net/) for drawing and analysing DAGs.

4. Causal Inference II: Estimation

4.1. Learning objectives

By the end of this chapter you should be able to:

- Estimate the propensity score and use it to construct matched cohorts, weighted samples, and stratified estimates.
- Implement inverse-probability-of-treatment weighting (IPTW) and recognise its sensitivity to extreme weights.
- Implement g-computation and double-robust estimators (AIPW, TMLE) and explain why double robustness matters.
- Apply instrumental-variable estimation when an IV is available, and recognise its limits.
- Conduct sensitivity analyses (E-value, tipping-point) for unmeasured confounding.

4.2. Orientation

Chapter 3 specified what we want to estimate (potential-outcome contrasts) and the assumptions under which they are identified (exchangeability, positivity, consistency). This chapter covers how to compute the estimates. Five families of estimators dominate applied work: propensity-score matching, IPTW, g-computation, doubly-robust estimators (AIPW, TMLE), and instrumental variables.

The choice between estimators is partly a matter of taste and partly a matter of which assumptions you are willing to defend. IPTW is sensitive to positivity violations; g-computation requires correctly specified outcome models; doubly-robust estimators are robust to mis-specification of one

of the two but require both. The chapter covers each, illustrates with a worked example, and recommends when to use which.

4.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Pick the estimator deliberately. The default is to use whatever estimator is most familiar (typically logistic regression with covariates, which is g-computation). The right choice depends on which mis-specifications are likeliest in your data: if the outcome model is hard to specify (rare events, non-linear treatment effects), use IPTW or doubly robust; if positivity is weak (some covariate strata have near-zero treatment probability), use g-computation or doubly robust; if neither works, reconsider whether the question is identifiable from this data.

(Judgement 2.) The propensity score is a means, not an end. The point of the propensity score is to remove confounding, not to predict treatment. A propensity model that perfectly predicts treatment is a positivity disaster; an over-fitted model is worse than an under-fitted one. The biostatistician judges propensity models on the balance they produce, not on AUC or calibration of the propensity prediction.

(Judgement 3.) Sensitivity analysis is part of the result, not after it. Every causal estimate is paired with a sensitivity analysis: an E-value, a tipping point, a negative-control check. The biostatistician designs the sensitivity analysis when designing the primary analysis, not after the primary result is in.

These judgements distinguish causal estimation that informs decisions from regression with extra steps.

4.4. The propensity score

The **propensity score** is the conditional probability of treatment given covariates:

$$e(X) = \Pr(A = 1 \mid X).$$

Rosenbaum and Rubin (Rosenbaum & Rubin, 1983) showed that conditioning on $e(X)$ is sufficient to remove confounding under exchangeability. The propensity score is a one-dimensional summary of a high-dimensional covariate vector — easier to use for matching, weighting, and balance assessment.

Estimation: typically a logistic regression of A on X. Modern alternatives use machine-learning models (gradient boosting, random forests) for the propensity estimation step. The choice has implications for double-robust estimators (more later).

```
library(MatchIt)

ps_model <- glm(treatment ~ age + sex + bmi + comorb,
                data = cohort, family = binomial)
cohort$ps <- predict(ps_model, type = "response")

summary(cohort$ps)
#>     Min.  1st Qu.  Median   Mean  3rd Qu.    Max.
#>   0.082    0.231   0.401  0.412   0.587    0.918
```

Check positivity: are propensity scores reasonably distributed in both treatment groups?

```
ggplot(cohort, aes(x = ps, fill = factor(treatment))) +
  geom_histogram(position = "dodge", binwidth = 0.05) +
  labs(x = "Propensity score", fill = "Treatment")
```

If the histograms are far apart, positivity is in trouble. If many treated patients have ps near 1 or many untreated have ps near 0, those observations carry little information for the comparison.

4.5. Propensity-score matching

Match each treated patient to one or more untreated patients with similar propensity scores; analyse the matched cohort.

```r
library(MatchIt)

m <- matchit(treatment ~ age + sex + bmi + comorb,
             data = cohort, method = "nearest",
             ratio = 1)
matched <- match.data(m)

summary(m)
#> [balance table]
```

Check balance: standardised mean differences should be under 0.10 for all confounders (a standard threshold).

Then estimate on the matched cohort:

```r
fit <- lm(outcome ~ treatment, data = matched)
summary(fit)
```

Propensity matching estimates the **ATT** by default (the effect among the treated). To estimate the **ATE**, match treated to untreated and untreated to treated (more complex; see `MatchIt` documentation).

4.6. Inverse-probability weighting (IPTW)

Weight each observation by the inverse of the probability of receiving the treatment they actually received:

$$w_i = \frac{A_i}{e(X_i)} + \frac{1 - A_i}{1 - e(X_i)}.$$

Treated patients with low propensity get high weight (rare in the data, given high weight to compensate); similarly for untreated.

Then fit an outcome model with the weights:

```
cohort$w <- ifelse(cohort$treatment == 1,
                   1/cohort$ps,
                   1/(1-cohort$ps))

library(survey)
des <- svydesign(~1, weights = ~w, data = cohort)
fit <- svyglm(outcome ~ treatment, design = des)
summary(fit)
```

`svyglm` produces robust standard errors that account for the weighting.

IPTW estimates the **ATE**. Stabilised weights (multiply by the marginal probability of the observed treatment) reduce the variance:

$$w_i^{\text{stab}} = \frac{A_i \Pr(A=1)}{e(X_i)} + \frac{(1-A_i)\Pr(A=0)}{1-e(X_i)}.$$

The major weakness: extreme weights. If a few treated patients have ps near 0.05, their weight is 20; their contribution dominates the estimate. **Trim** weights at some quantile (say, the 1st and 99th percentile) or truncate at a fixed value.

4.7. g-computation

g-computation (J. Robins, 1986) is the direct application of the identification formula from Chapter 3:

1. Fit an outcome model $E[Y \mid A, X]$ on the data.
2. Predict $E[Y \mid A=1, X_i]$ for every observation i (including untreated ones); average to get $E[Y(1)]$.
3. Predict $E[Y \mid A=0, X_i]$ for every observation; average to get $E[Y(0)]$.
4. Compute the difference (or ratio).

```r
fit <- glm(outcome ~ treatment + age + sex + bmi + comorb,
           data = cohort, family = binomial)

cohort$y1 <- predict(fit, newdata = transform(cohort,
                                              treatment = 1),
                     type = "response")
cohort$y0 <- predict(fit, newdata = transform(cohort,
                                              treatment = 0),
                     type = "response")

ate <- mean(cohort$y1) - mean(cohort$y0)
ate
#> [1] 0.073
```

Bootstrap for confidence intervals (the standard error in the regression output is for the conditional effect, not the marginal ATE).

g-computation requires correctly specifying the outcome model. Misspecification biases the result. With many covariates and complex interactions, this is hard.

4.8. Doubly-robust estimators

Doubly-robust estimators combine a propensity model and an outcome model in a way that gives consistent estimates if either one is correctly specified (Bang & Robins, 2005; J. M. Robins et al., 1994). The two main families:

AIPW (Augmented IPW) combines IPTW with an outcome adjustment:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_i \left[\frac{A_i(Y_i - \hat{m}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-A_i)(Y_i - \hat{m}_0(X_i))}{1-\hat{e}(X_i)} + \hat{m}_1(X_i) - \hat{m}_0(X_i)\right]$$

where $\hat{m}_a(X) = E[Y \mid \widehat{A = a}, X]$.

TMLE (Targeted Maximum Likelihood Estimation) (Laan & Rubin, 2006) is a more sophisticated double-robust estimator that achieves better finite-sample properties.

Both rely on **cross-fitting** (Chernozhukov et al., 2018) when machine-learning estimators are used for the nuisance functions; cross-fitting splits the data, fits the nuisance models on one half, computes the ATE contributions on the other, and averages. Without cross-fitting, ML-based double-robust estimators have poor coverage.

In R, the `tmle` and `WeightIt` + `marginaleffects` packages implement these. The `lmtp` package handles longitudinal extensions.

```
library(tmle)
result <- tmle(Y = cohort$outcome,
               A = cohort$treatment,
               W = cohort[, c("age","sex","bmi","comorb")])
summary(result)
```

Doubly-robust is the default recommendation for modern applied causal inference: it is robust to one model mis-specification, integrates with ML for nuisance estimation, and produces valid CIs under cross-fitting.

Check your understanding: when matching beats weighting

Question. A cohort study has 200 treated patients and 800 untreated. The propensity-score distribution shows substantial overlap, but a few treated patients have propensity scores near 0.05. Which estimator is the safer choice: IPTW or 1:1 propensity matching?
Answer.
1:1 matching. The treated patients with propensity near 0.05 would receive weights of 20 in IPTW, dominating the estimate and inflating the variance. Matching either pairs them with the small number of untreated patients with similar low propensity (limiting extrapolation) or excludes them (with explicit documentation). Either choice is defensible and transparent. IPTW also has principled trimming strategies but matching is more interpretable for an audience evaluating positivity violations directly. The trade is that 1:1 matching estimates the ATT (not the ATE), so the substantive question matters.

4.9. Instrumental variables

When unmeasured confounding is present and cannot be addressed by adjustment, an instrumental variable (IV) provides a different identification route. An IV Z satisfies:

- Z affects A.
- Z has no direct effect on Y except through A (the **exclusion restriction**).
- Z is unconfounded with Y.

Examples in clinical research: random assignment in an RCT (the canonical IV); physician prescribing preferences; geographic variation in availability.

Two-stage least squares estimates the **local average treatment effect** (LATE) — the effect among 'compliers', i.e., patients whose treatment status is affected by the instrument:

```
library(AER)
fit_iv <- ivreg(outcome ~ treatment + age + sex |
                ins_pref + age + sex,
                data = cohort)
summary(fit_iv)
```

The first half of the formula is the structural equation; the second half is the first-stage equation (with the IV `ins_pref` instead of `treatment`).

IV estimation requires strong instruments (`F-statistic > 10` in the first stage is a rule of thumb) and the exclusion restriction (rarely falsifiable from data). Sensitivity analysis under violation of exclusion is a active research area (Conley et al., 2012).

4.10. Sensitivity analysis: the E-value

The E-value (VanderWeele & Ding, 2017) quantifies how strong an unmeasured confounder would need to be to overturn an observed association. For an observed risk ratio RR_{obs}:

$$\text{E-value} = \text{RR}_{\text{obs}} + \sqrt{\text{RR}_{\text{obs}}(\text{RR}_{\text{obs}} - 1)}.$$

The E-value is the minimum strength of association, on the risk-ratio scale, that an unmeasured confounder would need to have with both the exposure and the outcome to fully explain the observed exposure-outcome association.

For $\text{RR}_{\text{obs}} = 2.0$, E-value $= 3.41$. Interpretation: an unmeasured confounder would need to be associated with both exposure and outcome with RR of at least 3.41 to overturn the observed association of 2.0. Whether 3.41 is plausible depends on the substantive context.

The `EValue` R package computes E-values for various estimands (RR, OR, HR, RD).

```
library(EValue)
evalues.RR(est = 2.0, lo = 1.6, hi = 2.5)
#>             E-value
#> RR    point     3.41
#> RR    CI bound  2.59
```

Report the E-value alongside every observational causal estimate.

4.11. Worked example: estimating the ATE of SGLT2 on

12-month mortality

Continuing the example from Chs 1-2.

```
library(tidyverse)
library(MatchIt)
library(survey)
library(EValue)

# 1. Propensity score
ps_model <- glm(sglt2 ~ age + sex + ef + ntpro_bnp +
                  egfr + dm + ckd + ace_arb,
                data = hf_cohort, family = binomial)
```

```r
hf_cohort$ps <- predict(ps_model, type = "response")

# Check positivity
ggplot(hf_cohort, aes(x = ps, fill = factor(sglt2))) +
  geom_histogram(position = "dodge", binwidth = 0.05)

# 2. IPTW (stabilised)
hf_cohort <- hf_cohort |>
  mutate(p_a = mean(sglt2),
         w = ifelse(sglt2 == 1,
                    p_a / ps,
                    (1 - p_a) / (1 - ps)),
         # trim at 1st and 99th percentile
         w = pmin(pmax(w, quantile(w, 0.01)),
                  quantile(w, 0.99)))

# Check balance after weighting
# (use cobalt::bal.tab() in production)

# 3. IPTW estimate
des <- svydesign(~1, weights = ~w, data = hf_cohort)
fit_iptw <- svyglm(mort_12mo ~ sglt2, design = des,
                   family = binomial)
summary(fit_iptw)
# log-OR for sglt2: ...

# 4. g-computation for cross-check
fit_g <- glm(mort_12mo ~ sglt2 + age + sex + ef +
               ntpro_bnp + egfr + dm + ckd + ace_arb,
             data = hf_cohort, family = binomial)
y1 <- mean(predict(fit_g, transform(hf_cohort, sglt2 = 1),
                   type = "response"))
y0 <- mean(predict(fit_g, transform(hf_cohort, sglt2 = 0),
                   type = "response"))
rr <- y1 / y0
rr
#> [1] 0.78
```

```
# 5. E-value
evalues.RR(rr)
#>             E-value
#> RR    point    1.87
```

The IPTW and g-computation point estimates agree (the same data, two estimation routes); the discrepancies between them — if any — flag misspecification. The E-value of 1.87 means that an unmeasured confounder would need to be associated with both SGLT2 use and mortality with RR 1.87 to explain the 22% mortality reduction observed; whether that is plausible is a substantive judgement.

A doubly-robust analysis (TMLE or AIPW) provides the recommended primary estimate; IPTW and g-computation provide cross-checks.

4.12. Collaborating with an LLM on causal-inference estimation

Three patterns.

Prompt 1: 'Implement IPTW for this dataset.' Provide the data structure and the confounders.

What to watch for. The LLM produces working IPTW code. It typically uses unstabilised weights, does not trim, and does not check balance. Push for the production version: stabilised, trimmed, with balance diagnostics.

Verification. Check the output against a textbook formula. Plot the propensity-score distribution and inspect balance with `cobalt::bal.tab()`.

Prompt 2: 'Walk me through whether this analysis is robust to unmeasured confounding.' Provide the analysis and the data sources.

What to watch for. The LLM gives a generic discussion of E-values. Push for specifics: which unmeasured confounders are most likely, how strong they would need to be to overturn the result.

Verification. Compute the E-value yourself; it is one formula. The interpretive judgement (is the required confounding strength plausible?) is yours.

Prompt 3: 'Compare the ATE and ATT for this analysis.' Provide both estimates.

What to watch for. The LLM correctly distinguishes the two and explains when each applies. It tends to miss the practical implication: which one matches the decision the report informs.

Verification. Map the estimand to the decision explicitly. Different audiences need different estimands.

The meta-pattern: LLMs are useful for the syntactic mechanics (writing the IPTW or g-computation code) and weak at the substantive judgement (which estimator, which confounders, which E-value to report). Use the LLM for the code; bring the substantive reasoning yourself.

4.13. Principle in use

Three habits.

1. **Match estimator to the assumption you can defend.** IPTW requires positivity. g-computation requires outcome-model specification. Doubly-robust requires one of the two. Choose by what you can defend.
2. **Cross-check estimators.** Run two estimators that make different mis-specifications. Agreement is evidence; disagreement flags a problem.
3. **Report sensitivity alongside the estimate.** The E-value (or equivalent) is part of the result, not a paragraph at the end of the discussion.

4.14. Exercises

1. Take a public observational dataset (e.g., NHANES subset) and compute the ATE using IPTW for a binary treatment and a binary

outcome of your choice. Report the propensity-score distribution, the balance table, and the trimmed estimate.

2. Repeat the analysis with g-computation. Compare the point estimates and confidence intervals.
3. Compute the E-value for your estimate. Discuss whether unmeasured confounding of the required strength is plausible.
4. Find a published observational study with a reported risk ratio. Compute the E-value for that estimate. Read the discussion for the authors' handling of unmeasured confounding; would you reach the same conclusion?
5. Extend the SGLT2 worked example to estimate the ATT (effect among initiators) using 1:1 propensity-score matching. Compare to the IPTW estimate and explain the difference.

4.15. Further reading

- Hernán & Robins (2020), *Causal Inference: What If*. The open-access textbook covering all the estimators in this chapter.
- Rosenbaum & Rubin (1983), 'The central role of the propensity score in observational studies for causal effects'. The foundational paper for propensity-score methods.
- VanderWeele & Ding (2017), 'Sensitivity analysis in observational research: introducing the E-value'. The reference for the E-value.
- The `MatchIt`, `WeightIt`, `tmle`, and `marginaleffects` R packages are the modern toolkit; their vignettes are the practical references.

5. Mediation Analysis

5.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish total, direct, and indirect effects in the potential-outcomes mediation framework.
- Apply the Baron-Kenny approach to a simple mediation question and recognise its limitations.
- Implement counterfactual mediation analysis for natural direct and indirect effects (NDE/NIE) using the `mediation` and `CMAverse` R packages.
- Reason about the four identifying assumptions for mediation and the conditions under which mediation effects are identifiable.
- Conduct sensitivity analysis for unmeasured mediator-outcome confounding.

5.2. Orientation

Mediation analysis decomposes a total causal effect into the part that flows through a specified mechanism (the indirect effect) and the part that does not (the direct effect). 'Does treatment X reduce mortality because it lowers blood pressure, or for some other reason?' is a mediation question. The answer requires both causal-inference machinery (Chs 3-4) and specific identifying assumptions about the mediator-outcome relationship.

The chapter develops three threads. **The framework**: total, direct, indirect effects in potential-outcomes notation, with the modern counterfactual definition. **The methods**: Baron-Kenny as a useful starting point, modern

counterfactual mediation as the production tool. **The assumptions**: four identifying conditions that must be defended for mediation to be valid.

The framing inherits the causal-inference discipline from Chapter 3. Mediation is a special case of causal inference where the analyst wants to decompose the effect; it inherits all the assumptions of basic causal inference plus additional ones about the mediator.

5.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Mediation requires more assumptions, not fewer. A causal estimate of the total effect requires no unmeasured exposure-outcome confounding (Ch 3). A causal estimate of the indirect effect through a mediator additionally requires no unmeasured mediator-outcome confounding, no exposure-mediator- outcome confounding interaction, and (for some estimators) no mediator-outcome interaction in the target population. Each new assumption is a substantive claim. The biostatistician does not run mediation analysis on data where the basic causal analysis is shaky.

(Judgement 2.) Pick the right effect for the question. Total effect, controlled direct effect, natural direct effect, natural indirect effect — each is a different counterfactual contrast and answers a different question. The total effect is sometimes right; the natural direct/indirect decomposition is right when the question is about mechanism; the controlled direct effect is right when the question is about an intervention that fixes the mediator.

(Judgement 3.) Sensitivity analysis is mandatory. The unmeasured mediator-outcome confounding assumption is rarely defensible without quantification. Modern mediation packages (`CMAverse`, `mediation`) provide sensitivity analyses; report them.

These judgements distinguish a mediation analysis that informs mechanistic claims from regression with a mediator term added.

5.4. The framework

Notation:

- A: exposure or treatment.
- M: mediator (a post-exposure, pre-outcome variable thought to lie on the causal path from A to Y).
- Y: outcome.
- X: pre-exposure confounders.
- $Y(a, m)$: potential outcome under exposure a and mediator value m.

The basic decomposition:

Total effect (TE):

$$\text{TE} = E[Y(1) - Y(0)]$$

the effect of exposure on outcome regardless of mechanism.

Controlled direct effect (CDE) at mediator level m:

$$\text{CDE}(m) = E[Y(1, m) - Y(0, m)]$$

the effect of exposure when the mediator is held fixed at m. Useful when the question is about an intervention that controls the mediator.

Natural direct effect (NDE):

$$\text{NDE} = E[Y(1, M(0)) - Y(0, M(0))]$$

the effect of exposure on outcome when the mediator is fixed at the level it would have taken under the control condition. The 'direct' effect that bypasses the mediator.

Natural indirect effect (NIE):

$$\text{NIE} = E[Y(1, M(1)) - Y(1, M(0))]$$

the effect on outcome of changing the mediator from its value under control to its value under exposure, holding exposure fixed at 1. The 'indirect' effect through the mediator.

The decomposition: TE = NDE + NIE. (Note that the intuition 'TE = direct + indirect' holds for natural effects but not for controlled direct effects, where the decomposition is more complex.)

The 'natural' framing matches the question 'how much of the total effect goes through this mechanism' more naturally than the 'controlled' framing matches it. The trade-off: natural effects require an additional identifying assumption that controlled effects do not.

5.5. Identifying assumptions

Four assumptions for identification of natural mediation effects (VanderWeele, 2015):

1. **No unmeasured exposure-outcome confounding** given X.
2. **No unmeasured mediator-outcome confounding** given X (and the exposure).
3. **No unmeasured exposure-mediator confounding** given X.
4. **No exposure-induced mediator-outcome confounding** (sometimes called the 'cross-world' assumption) — technically, no mediator-outcome confounder that is itself caused by the exposure.

Assumptions 1-3 are extensions of the basic exchangeability assumption from Ch 3. Assumption 4 is specific to natural-effect identification and is the hardest to defend; it cannot be guaranteed by design in observational data and rarely in trial data either. The controlled direct effect requires only 1-3.

5.6. Baron-Kenny

The classic approach (Baron & Kenny, 1986): regress Y on A to get the total effect c; regress M on A to get the $A \to M$ path a; regress Y on A and M to get the $M \to Y$ path b and the direct effect c'. Indirect effect $= a \cdot b$; total = direct + indirect $= c' + a \cdot b = c$ (under linear models with no interaction).

```
fit_y_a   <- lm(y ~ a + x, data = d)          # total
fit_m_a   <- lm(m ~ a + x, data = d)          # path a
fit_y_am  <- lm(y ~ a + m + x, data = d)      # paths c', b
```

```
c  <- coef(fit_y_a)["a"]
a_path <- coef(fit_m_a)["a"]
b_path <- coef(fit_y_am)["m"]
c_prime <- coef(fit_y_am)["a"]

c                          # total
c_prime                    # direct
a_path * b_path            # indirect
c_prime + a_path * b_path  # should = c
```

Baron-Kenny works when:

- All variables are continuous.
- The models are linear with no interactions (in particular, no $A \times M$ interaction in the outcome model).
- All four mediation assumptions hold.

It breaks (or biases) when:

- The outcome is binary (logistic regression's coefficients do not decompose simply).
- The exposure and mediator interact in their effect on the outcome.
- Exposure-induced mediator-outcome confounders are present.

For modern applied mediation, Baron-Kenny is a useful first-pass diagnostic but not the right tool for publication.

5.7. Counterfactual mediation analysis

The modern approach uses potential outcomes directly to define and estimate NDE and NIE. The `mediation` R package (Imai et al., 2010) and the `CMAverse` package (Shi et al., 2021) implement this.

```
library(mediation)

# the mediator model
fit_m <- lm(m ~ a + x, data = d)
```

```
# the outcome model (allowing A:M interaction)
fit_y <- lm(y ~ a * m + x, data = d)

# mediation analysis
med_result <- mediate(fit_m, fit_y,
                      treat = "a", mediator = "m",
                      sims = 1000)
summary(med_result)
#> Causal Mediation Analysis
#>
#> Quasi-Bayesian Confidence Intervals
#>
#>                Estimate  95% CI Lower  95% CI Upper p-value
#> ACME (control)    0.043         0.018         0.072  <2e-16
#> ACME (treated)    0.049         0.022         0.080  <2e-16
#> ADE (control)     0.082         0.035         0.131  <2e-16
#> ADE (treated)     0.088         0.039         0.140  <2e-16
#> Total Effect      0.131         0.083         0.182  <2e-16
#> Prop. Mediated   0.339         0.193         0.521  <2e-16
```

The output:

- ACME (Average Causal Mediation Effect) is the natural indirect effect.
- ADE (Average Direct Effect) is the natural direct effect.
- Total effect is the sum.
- Proportion mediated is the ratio of NIE to TE.

The 'control' and 'treated' versions reflect the $A \times M$ interaction; if the interaction is small, they agree.

For binary outcomes or non-linear models, `CMAverse` provides more flexibility:

```
library(CMAverse)

result <- cmest(data = d, model = "rb",
                outcome = "y", exposure = "a",
```

```
                mediator = "m",
                basec = c("x1", "x2"),
                yreg = "logistic",
                mreg = list("linear"),
                EMint = TRUE,
                astar = 0, a = 1, mval = list(0))
summary(result)
```

`CMAverse` supports linear, logistic, Poisson, multinomial, ordinal, time-to-event, and survival outcomes; multiple mediators (sequential and joint); and effect modification. The `EMint = TRUE` includes the exposure-mediator interaction.

Check your understanding: when proportion mediated misleads

Question. A mediation analysis reports: TE = 0.05, NIE = 0.03, NDE = 0.02, proportion mediated = 60%. The total effect is small but a large fraction goes through the mediator. Is the 60% figure informative?

Answer.

It is mathematically correct but pragmatically misleading. Proportion mediated is the ratio of two estimated quantities; both are noisy. When TE is small, proportion mediated is unstable: a small change in TE or NIE produces a large swing in the ratio. The useful reporting is NIE and NDE in their natural units (with CIs), not the ratio. Reserve proportion mediated for cases where TE is well-estimated and substantial. Several published mediation analyses have reported proportion mediated of 80% or 100% when TE is near zero, producing misleading conclusions.

5.8. Sensitivity analysis

The mediator-outcome no-confounding assumption is typically the weakest. Sensitivity analysis quantifies how strong unmeasured mediator-outcome confounding would need to be to overturn the conclusion.

The `mediation` package's sensitivity:

```
sens <- medsens(med_result, sims = 500)
summary(sens)
plot(sens)
```

The output: at what level of mediator-outcome confounding (parameterised by ρ, the correlation between residuals in the mediator and outcome models) would the NIE drop to zero. A small ρ means the result is sensitive to even mild unmeasured confounding; a large ρ means substantial confounding would be required.

`CMAverse` provides analogous sensitivity analyses for its broader range of models.

5.9. Multiple and sequential mediators

Real questions often involve multiple potential mediators. Treatment X reduces CV mortality; the mechanism could be through blood pressure, lipids, glucose, or weight. Several extensions handle this:

Joint mediation treats the mediators as a vector; the joint NIE captures the indirect effect through the whole vector. Useful when the mediators are correlated and one-at-a-time analyses double-count.

Sequential mediation specifies a temporal order among the mediators (e.g., treatment → BP → lipids → outcome) and decomposes the indirect effect into the contribution of each.

Both are implemented in `CMAverse` and the `gformula` package. The complexity rises quickly; the assumptions multiply. Reserve multiple-mediator analyses for questions where the substantive interest justifies the methodological burden.

5.10. Worked example: does SGLT2 reduce mortality through

blood pressure?

Continuing the SGLT2 example. The substantive question: of the observed 22% mortality reduction under SGLT2 vs. no SGLT2, how much goes through the known mechanism of lowered blood pressure?

```
library(mediation)

# mediator model
fit_m <- lm(sbp_3mo ~ sglt2 + age + sex + ef +
              egfr + baseline_sbp,
            data = hf_cohort)

# outcome model
fit_y <- glm(mort_12mo ~ sglt2 * sbp_3mo + age + sex +
               ef + egfr + baseline_sbp,
             data = hf_cohort, family = binomial)

# mediation
med <- mediate(fit_m, fit_y,
               treat = "sglt2", mediator = "sbp_3mo",
               sims = 1000)
summary(med)
```

Suppose the output shows NIE = -0.04 (SGLT2 reduces mortality by 4 percentage points through SBP), NDE = -0.18 (SGLT2 reduces mortality by 18 percentage points not through SBP), TE = -0.22, proportion mediated = 18%.

Interpretation: blood-pressure reduction explains a small fraction of the SGLT2 mortality benefit; most of the benefit is through other mechanisms (likely osmotic diuresis and direct cardiac effects). This is the kind of substantive insight mediation analysis is designed to produce.

The sensitivity analysis (running `medsens(med)`) tells you how much unmeasured confounding of SBP-mortality would be required to attenuate the NIE to zero.

5.11. Collaborating with an LLM on mediation analysis

Three patterns.

Prompt 1: 'Write the mediation model for this question.' Provide the exposure, mediator, outcome, and confounders.

What to watch for. The LLM produces working code using `mediation` or `CMAverse`. It frequently omits the `EMint = TRUE` (exposure-mediator interaction), which produces biased natural effects when the interaction is real. Push back: 'should we include exposure-mediator interaction?'

Verification. Read the resulting code; confirm the interaction is included if appropriate. Run a sensitivity analysis.

Prompt 2: 'Interpret this mediation output for a non-statistical audience.' Provide the output.

What to watch for. The LLM produces clean prose for the NIE/NDE distinction. It tends to over-emphasise proportion mediated. Push for a sentence on the decomposition in absolute units rather than the ratio.

Verification. The interpretation should match the estimands. NIE in absolute units is the most interpretable quantity for most audiences.

Prompt 3: 'What are the assumptions for this mediation analysis to be valid?' Provide the analysis.

What to watch for. The LLM lists the four assumptions. It is generally vague about the cross-world assumption. Push for specifics: 'what exposure-induced mediator-outcome confounders are plausible?'

Verification. Compare the LLM's list to the VanderWeele (2015) reference. The substantive judgement (which assumptions are plausible) is yours.

The meta-pattern: LLMs are good at the syntactic mechanics (the right R function, the standard output) and weak at the substantive judgements (which mediation effect to estimate, which assumptions are plausible). Use them for code, bring substantive reasoning yourself.

5.12. Principle in use

Three habits.

1. **Run a basic causal analysis first.** If the total-effect causal analysis is shaky, the mediation analysis cannot rescue it. Mediation inherits all the assumptions of basic causal inference plus more.
2. **Include the exposure-mediator interaction.** Without it, NDE and NIE may be biased. Include it by default; remove it only after testing.
3. **Report sensitivity analysis for mediator-outcome confounding.** The assumption is rarely defensible without quantification.

5.13. Exercises

1. For a research question of your choice with a plausible mediator, write the mediation framework: exposure, mediator, outcome, confounders for each relationship, the four mediation assumptions.
2. Implement Baron-Kenny on simulated data with no exposure-mediator interaction. Verify $c = c' + ab$ holds.
3. Repeat exercise 2 with an exposure-mediator interaction. Note where Baron-Kenny breaks and compare to the counterfactual analysis from `mediation`.
4. Run a sensitivity analysis for unmeasured mediator-outcome confounding on a published mediation result. Comment on whether the level of confounding required to overturn the result is plausible.
5. Read a published mediation analysis. Identify the four mediation assumptions and the authors' defence (or absence of defence) of each. Propose improvements.

5.14. Further reading

- VanderWeele (2015), *Explanation in Causal Inference*. The reference textbook for modern mediation analysis.
- Imai et al. (2010), 'A general approach to causal mediation analysis'. The methods paper for the `mediation` package.
- Shi et al. (2021), 'CMAverse: a suite of functions for reproducible causal mediation analyses'. The current applied tool.
- The `mediation` and `CMAverse` package documentation are the practical references.

Part III.

Correlated Data and Time-to-Event

6. Longitudinal and Correlated Data, Applied

6.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish marginal models (GEE) from conditional models (LMM, GLMM) and choose between them based on the question and the structure of the correlation.
- Fit linear mixed models with `lme4::lmer()` and generalised linear mixed models with `lme4::glmer()` or `glmmTMB::glmmTMB()`, and interpret fixed and random effects.
- Implement generalised estimating equations (GEE) with `geepack::geeglm()` and select an appropriate working correlation structure.
- Diagnose convergence failures and singular fits in mixed models, and apply remediations.
- Joint-model longitudinal and time-to-event outcomes when the question requires it.

6.2. Orientation

Most biomedical data has correlation structure: repeated measures on the same patient, multiple patients within the same hospital or cluster, observations close in time more similar than observations far apart. Ignoring the correlation produces invalid standard errors and sometimes biased point estimates. The introductory volume covered linear mixed models as a model class; this chapter develops them at applied depth — model- building,

diagnostics, when to use marginal vs. conditional, and the interaction with missing data.

The chapter is organised around the central choice: marginal vs. conditional models. Marginal models (GEE) target population-average effects. Conditional models (LMM, GLMM) target subject-specific effects with explicit random effects. The two are not interchangeable; choosing wrongly produces the wrong answer to the question.

6.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Marginal vs. conditional matches the question. A 'population-average effect' is the average effect across all patients in the population. A 'subject-specific effect' is the effect for a typical patient with a specific random-effect value. For linear models the two coincide; for non-linear models (logistic, Poisson, survival) they differ. The biostatistician picks the framework based on the question and reports the choice transparently.

(Judgement 2.) The random-effects structure is a substantive claim. Adding random slopes vs. intercepts, or specifying which variables get random effects, encodes assumptions about which sources of variation matter. Maximum-likelihood estimation will fit any structure you specify; only domain reasoning informs which structure is right. The biostatistician justifies the random-effects structure in the methods, not by goodness-of-fit alone.

(Judgement 3.) Missing data interacts with correlation. Mixed models handle missing data under the missing-at-random (MAR) assumption when the model is correctly specified for the longitudinal trajectory and the covariates predicting missingness. GEE requires missing-completely-at-random (MCAR) for unbiased estimation unless inverse-probability-of-censoring weighting is used. The biostatistician chooses the framework with attention to the missing- data mechanism (Ch 10).

6.4. Marginal vs. conditional: the central distinction

Consider repeated measures of blood pressure on hypertensive patients receiving treatment X.

Conditional (subject-specific) model:

$$\text{SBP}_{ij} = \beta_0 + \beta_1 \text{Tx}_{ij} + u_i + \epsilon_{ij}, \quad u_i \sim N(0, \tau^2).$$

β_1 is the within-patient effect of treatment for a typical patient. The random intercept u_i captures each patient's baseline level.

Marginal (population-average) model:

$$E[\text{SBP}_{ij}] = \beta_0^* + \beta_1^* \text{Tx}_{ij}$$

with a working correlation structure for the within- patient correlation. β_1^* is the average effect across the population.

For continuous outcomes with linear link, $\beta_1 = \beta_1^*$. For binary outcomes with logit link, the two differ. Concretely: the conditional log-odds of hypertension under treatment vs. control is not the same as the population log-odds-ratio; the conditional estimate is generally larger in magnitude.

When does each apply?

Conditional models when:

- The question is about within-patient or subject-specific effects ('does the treatment lower THIS patient's BP').
- The random-effects structure is a substantive part of the model (e.g., centre-level random effects capturing centre-specific variation).
- Missing data is MAR rather than MCAR.

Marginal models when:

- The question is about population-average effects ('what is the average BP reduction across the population on this treatment').
- Subject-specific random effects are nuisance.
- The correlation structure can be approximated by a working correlation that does not need to be exactly correct.

6.5. Linear mixed models with `lme4`

The standard tool:

```
library(lme4)

fit <- lmer(sbp ~ visit * treatment + age + sex +
              (1 + visit | id),
            data = bp_data)
summary(fit)
```

Reading the formula:

- `sbp ~ visit * treatment + age + sex`: fixed effects. Visit and treatment with their interaction; baseline age and sex.
- `(1 + visit | id)`: random intercept and random slope on visit, by patient ID.

Reading the output:

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 142.3    11.93
          visit         3.2     1.79    0.21
 Residual              45.7     6.76

Fixed effects:
                     Estimate Std. Error t value
(Intercept)           132.45      2.13     62.2
visit                  -1.42      0.34     -4.2
treatment              -3.71      1.45     -2.6
visit:treatment        -2.11      0.45     -4.7
age                     0.18      0.04      4.5
sex                    -2.05      1.83     -1.1
```

The fixed effects are the population-typical coefficients. The random-effect variances tell you how much patients differ from each other in baseline level and slope.

Confidence intervals via `confint()`:

```
confint(fit, method = "profile")
```

The default `lmer` does not produce p-values for fixed effects (a longstanding deliberate choice). For p-values, use `lmerTest::lmer()` (Satterthwaite degrees of freedom) or `parameters::model_parameters()`. Both are reasonable; cite which you used.

6.5.1. Random-effects structure

The `(1 + visit | id)` notation specifies:

- A random intercept (each patient has their own baseline).
- A random slope on visit (each patient has their own trajectory).
- The correlation between intercept and slope is estimated.

Other patterns:

- `(1 | id)`: random intercept only.
- `(0 + visit | id)`: random slope only, no random intercept (rarely the right choice).
- `(1 | id) + (1 | center)`: nested random effects (patients nested in centres).
- `(1 | center/id)`: same as above with shorthand.
- `(1 | center) + (1 | id)`: crossed random effects.

The random-effects structure should match the data's hierarchy. Patient-within-centre is nested; treatment- within-patient (where treatments cycle through patients in a crossover) is crossed.

6.5.2. Convergence and singular fits

`lmer` sometimes warns about singular fits ('boundary (singular) fit: see ?isSingular'). This means a variance component is estimated at zero — the model believes there is no patient-level variation. Causes:

- Small data with few patients per cluster.

- Random-effects structure that is not supported by the data.
- Genuinely no variation at the level specified.

Responses:

- Simplify the random-effects structure (drop a random slope, keep the intercept).
- Use `glmmTMB` (more robust optimiser) or `brms` (Bayesian, no convergence pathologies of the same kind).
- Accept the singular fit and report it.

The `lme4` package's troubleshooting vignette (`?lme4::convergence`) is the practical reference.

6.6. Generalised linear mixed models

For non-Gaussian outcomes, use `glmer` or `glmmTMB`:

```
library(glmmTMB)

fit_glmm <- glmmTMB(diabetic ~ age + bmi + treatment +
                      (1 | center),
                    data = patients,
                    family = binomial)
summary(fit_glmm)
```

The interpretation of fixed effects in a logistic GLMM is **conditional on the random effect**. The fixed- effect coefficient on treatment is the log-odds change for a patient with their centre's typical random effect. The marginal (population-average) odds ratio differs.

`glmmTMB` is the modern recommendation for GLMMs in R. It is faster and more robust than `glmer`, supports a wider range of distributions (zero-inflated, beta, ordinal), and produces cleaner output. `lme4::glmer` remains the reference but `glmmTMB` is the practical default.

6.7. GEE

`geepack` is the standard implementation:

```
library(geepack)

fit_gee <- geeglm(diabetic ~ age + bmi + treatment,
                  data = patients,
                  id = center,
                  family = binomial,
                  corstr = "exchangeable")
summary(fit_gee)
```

GEE specifies a marginal mean structure and a working correlation. The `corstr` choices:

- `"independence"`: assumes no within-cluster correlation. Robust standard errors fix the SEs even when correlation is present; estimator remains consistent.
- `"exchangeable"`: assumes constant correlation within cluster. Right for clusters with no internal ordering.
- `"ar1"`: AR(1) correlation (decreasing with time lag). Right for repeated measures over time with declining correlation.
- `"unstructured"`: estimates each pairwise correlation. Most flexible; needs sample size.

The point estimate's robust standard errors (from `summary(fit_gee)$coefficients`) are valid even if the working correlation is wrong. This is the key property: **GEE is robust to misspecification of the correlation structure**, making it a workhorse for applied work.

The trade: GEE estimates marginal effects, not conditional. If the question is about subject-specific effects, GEE is the wrong tool.

Check your understanding: when GEE is the right tool

Question. A multi-centre study has 30 centres and 2,000 patients. The outcome is binary; the question is 'on average across the population, what is the OR for treatment vs. control'. Should the analysis

use GEE or GLMM?
Answer.
GEE is the cleaner choice. The question asks about the population-average effect (marginal OR), which GEE estimates directly. A GLMM with random centre intercept estimates a conditional-on-centre OR; converting that to a marginal OR is doable but adds complexity. The GEE's robust SEs handle the within-centre correlation without requiring a correctly specified correlation structure. Reserve the GLMM for when the question is genuinely about subject-specific (or centre-specific) effects, or when the missing-data mechanism makes GEE biased.

6.8. Joint models for longitudinal and time-to-event

A common applied question: a longitudinal biomarker trajectory is associated with time to an event (death, disease progression). The naive analysis (a Cox regression with the biomarker as a time-varying covariate) is biased when the biomarker is measured with error or measured intermittently. **Joint models** fit a longitudinal model for the biomarker and a time-to-event model for the event simultaneously, linked by shared random effects.

The `JM` and `JMbayes2` packages implement this in R:

```
library(JM)

# longitudinal submodel
fit_long <- lme(biomarker ~ time + treatment,
                random = ~ time | id, data = long_data)

# survival submodel
fit_surv <- coxph(Surv(time, event) ~ treatment,
                  data = baseline_data, x = TRUE)

# joint
fit_joint <- jointModel(fit_long, fit_surv,
```

```
                          timeVar = "time")
summary(fit_joint)
```

The joint model produces:

- The longitudinal trajectory parameters with proper SEs.
- The hazard ratio for the biomarker as a time-varying covariate, accounting for measurement error.
- The 'association parameter' linking the two submodels.

Joint models are computationally heavier and require careful diagnostics; reserve them for when the question genuinely demands the linkage. For most applied work, separate longitudinal and survival analyses are adequate.

6.9. Missing data in longitudinal analysis

Longitudinal data is almost always missing some visits. The mechanism matters.

MCAR: missingness independent of all observed and unobserved data. Both LMM and GEE are unbiased.

MAR: missingness depends on observed data. LMM is unbiased under correct model specification; GEE is biased without inverse-probability-of-censoring weighting (IPCW).

MNAR: missingness depends on unobserved data. Both biased; sensitivity analyses required (Ch 10).

The implication: prefer LMM (or GLMM) over GEE when missing data is plausibly MAR rather than MCAR. The trade-off with the marginal-vs.-conditional question is real.

6.10. Worked example: a 12-month BP trajectory analysis

A trial randomised 200 hypertensive patients to treatment X or control; SBP measured at baseline, 3, 6, 9, 12 months. Some patients missed visits.

```
library(tidyverse)
library(lme4)
library(lmerTest)

bp <- read_csv("data/bp-trial.csv")
glimpse(bp)

# baseline characteristics
bp |>
  filter(visit == 0) |>
  group_by(treatment) |>
  summarise(n = n(),
            mean_age = mean(age),
            mean_sbp = mean(sbp))

# longitudinal model
fit <- lmer(sbp ~ visit * treatment + age + sex +
              (1 + visit | id),
            data = bp)
summary(fit)
anova(fit)   # via lmerTest, gives p-values

# random-effects diagnostics
ranef_summary <- ranef(fit)$id
plot(ranef_summary)        # check normality

# residual diagnostics
plot(fit)                  # fitted vs. residuals
qqnorm(resid(fit))         # residual QQ

# the contrast we care about: difference in slopes
```

```
emm <- emmeans::emmeans(fit, "treatment",
                        by = "visit",
                        at = list(visit = 12))
emm
pairs(emm)
```

The `emmeans` package extracts the contrast at month 12 between treatment and control, with appropriate SEs.

The reported result might be: at month 12, mean SBP in the treatment group is 11.3 mmHg (95% CI 8.4-14.2, $p < 0.001$) lower than control, adjusting for baseline characteristics. The model accounts for the correlation across visits within each patient.

The sensitivity analysis: refit under MNAR assumptions (Ch 10's pattern-mixture or selection models) to check whether the conclusion is robust to the missing-data mechanism.

6.11. Collaborating with an LLM on longitudinal data analysis

Three patterns.

Prompt 1: 'Build the mixed model for this longitudinal question.' Provide the data structure and the question.

What to watch for. The LLM produces a working `lmer` or `glmmTMB` formula. It often gets the random-effects structure wrong: omitting random slopes when the trajectory varies, including random intercepts on nested levels without justification. Push back on the random-effects choice.

Verification. The random-effects structure should match the data's hierarchy and the substantive question. Run with and without random slopes; compare.

Prompt 2: 'Should I use GEE or GLMM here?' Provide the question, the data, and the missing-data context.

What to watch for. The LLM gives a competent marginal-vs-conditional distinction. It tends to default to GLMM. Push for the missing-data implications (GEE biased under MAR, etc.).

Verification. The choice depends on the question (population-average vs. subject-specific) and the missing-data mechanism. Both factors must be considered.

Prompt 3: 'Diagnose this convergence failure in `lmer`.' Provide the warning and the model specification.

What to watch for. The LLM commonly suggests simplifying the random-effects structure or switching to `bobyqa` optimiser. These often work but sometimes mask a substantive issue (the random slope genuinely is zero, suggesting the trajectory does not vary by patient).

Verification. Try the LLM's suggestions and inspect the simpler model. If the simpler fit is well- behaved, the problem may have been over-specification.

The meta-pattern: LLMs are good for the syntactic mechanics (writing the formula, suggesting standard diagnostics) and weak at substantive judgement (which random effects belong, which framework matches the question). Use them for code, bring substantive reasoning yourself.

6.12. Principle in use

Three habits.

1. **Match marginal vs. conditional to the question.** The two estimate different parameters in non-linear models. Pick deliberately.
2. **Justify the random-effects structure.** The structure is a substantive claim; defend it in the methods.
3. **Address missing data explicitly.** Mixed models handle MAR under correct specification; GEE requires MCAR or IPCW. The choice interacts with the missing-data mechanism.

6.13. Exercises

1. Take a longitudinal dataset of your choice. Fit both an LMM and a GEE for the same outcome variable. Compare the point estimates and standard errors. Where they differ, explain why.
2. For a binary outcome with cluster sampling, fit a GLMM with random intercept and a GEE with exchangeable working correlation. Compare the conditional and marginal odds ratios; verify they differ.
3. Build a mixed model with a random slope on time. Inspect the patient-specific slopes (`ranef()`). Identify the patients with the most extreme trajectories.
4. For a longitudinal study with missing visits, conduct the primary LMM and a sensitivity analysis under MAR. (The MAR sensitivity analysis is the topic of Ch 10; for this exercise, use multiple imputation as a bridge.)
5. Read a published longitudinal analysis. Identify whether the framework is marginal or conditional, and whether the choice matches the stated question.

6.14. Further reading

- Fitzmaurice et al. (2011), *Applied Longitudinal Analysis* (2nd edition). The reference textbook.
- Diggle et al. (2002), *Analysis of Longitudinal Data* (2nd edition). The classical treatment.
- Rizopoulos (2012), *Joint Models for Longitudinal and Time-to-Event Data*. The reference for joint modelling.
- The `lme4`, `glmmTMB`, `geepack`, and `JM` package vignettes are the practical references.

7. Survival Analysis, Applied

7.1. Learning objectives

By the end of this chapter you should be able to:

- Construct Kaplan-Meier curves and read them correctly, including the censoring marks and the number-at-risk table.
- Fit a Cox proportional hazards model and check the proportional-hazards assumption with `cox.zph`.
- Distinguish cause-specific hazards from the Fine-Gray subdistribution hazard model when competing risks are present, and choose between them by question.
- Compute and interpret restricted mean survival time (RMST) and recognise when it is the right summary.
- Handle time-varying covariates with the (start, stop, event) data structure.
- Recognise and report immortal-time bias and conditioning-on-the-future errors.

7.2. Orientation

Survival analysis (more accurately, time-to-event analysis) is the methods area where biostatistics diverges most sharply from general statistics. Time-to- event data has censoring, the proportional-hazards assumption is testable, competing risks are common, and the most-reported summary (the hazard ratio) has known interpretive pitfalls. The introductory volume covered survival as a model class; this chapter develops the applied depth required for clinical research and regulatory submissions.

The chapter is organised in three threads. **Foundations**: censoring, survival functions, hazards, Kaplan-Meier. **Cox PH and beyond**: the workhorse model, proportional-hazards diagnostics, extensions for time-varying covariates and competing risks. **Modern alternatives**: RMST, parametric models when PH fails.

7.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) The hazard ratio is not always the right summary. A Cox HR of 0.7 is reported routinely; the substantive interpretation ('30% reduction in hazard') is widely misread as 'survival is 30% better'. When proportional hazards fails, the HR is a weighted average over follow-up time and may not summarise the effect at any specific time point. The biostatistician chooses the summary (HR, RMST, milestone survival, absolute risk difference) by what the audience needs to act on, not by software default.

(Judgement 2.) Competing risks change the question. A patient who dies of competing causes cannot subsequently experience the event of interest (e.g., disease recurrence). Treating death as 'censoring' in a standard Cox model overstates cumulative incidence. The biostatistician identifies competing risks and chooses cause-specific or subdistribution methods to match the question.

(Judgement 3.) Time-varying covariates are common sources of bias. The 'patient took drug X for at least 6 months' as a baseline covariate is immortal- time bias: patients who survived 6 months had to survive 6 months, so the comparison group is artificially impoverished. The biostatistician recognises immortal-time bias and reformulates the analysis using the (start, stop, event) structure or target-trial emulation.

7.4. Foundations: hazard, survival, Kaplan-Meier

For a time-to-event variable T:

Survival function $S(t) = \Pr(T > t)$. The probability of being event-free at time t.

Hazard function $\lambda(t)$. The instantaneous event rate among those still at risk:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}.$$

The two are related by:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\, du\right) = \exp(-\Lambda(t))$$

where $\Lambda(t)$ is the cumulative hazard.

Censoring is the defining feature: a patient is 'censored' at time c if their event time $T > c$ but T is not observed. The Kaplan-Meier estimator incorporates censoring:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of events at time t_i and n_i is the number at risk just before t_i.

```
library(survival)
library(survminer)

fit_km <- survfit(Surv(time, event) ~ treatment,
                  data = trial)
ggsurvplot(fit_km, data = trial,
           risk.table = TRUE,
           conf.int = TRUE,
           pval = TRUE)
```

The plot includes the survival curves, 95% CIs, the log-rank p-value, and (essential) the number-at-risk table. The number-at-risk table is non-negotiable in published KM plots; without it the reader cannot evaluate the precision of the curves at later time points.

7.5. The Cox proportional hazards model

The model:

$$\lambda(t \mid X) = \lambda_0(t) \exp(X^T \beta)$$

where $\lambda_0(t)$ is the (unspecified) baseline hazard. The coefficients β are log-hazard ratios.

Estimation via partial likelihood — no specification of $\lambda_0(t)$ required. This is what makes Cox the workhorse: minimal parametric assumptions, just proportional hazards.

```
fit_cox <- coxph(Surv(time, event) ~ treatment + age +
                   sex + stage,
                 data = trial)
summary(fit_cox)
#> ...
#>            coef exp(coef) se(coef)    z Pr(>|z|)
#> treatment -0.36      0.70     0.10 -3.6   0.0003
#> age        0.02      1.02     0.01  4.0   <0.001
#> ...
```

The `exp(coef)` is the hazard ratio. `treatment`'s HR of 0.70 means the hazard is 70% of the control hazard at any time, holding other covariates fixed.

7.5.1. Proportional-hazards assumption

The model assumes the hazard ratio is constant over follow-up. Test with `cox.zph`:

```
ph_test <- cox.zph(fit_cox)
print(ph_test)
plot(ph_test)
```

A significant p-value indicates non-proportional hazards. Inspection of the plot shows whether the violation is mild or severe.

If proportional hazards fails:

1. **Stratify** the violator: include it as a stratum variable (`strata(stage)`) rather than a covariate. The model fits separate baseline hazards within each stratum.
2. **Add a time-varying coefficient** for the violator (`tt(variable)` syntax in `coxph` or manually constructed time interactions).
3. **Switch summary**: report RMST, milestone survival, or hazard ratios at specific time intervals.

The first option is the most common.

7.5.2. Time-varying covariates

For covariates that change over follow-up (e.g., a biomarker measured periodically), use the (start, stop, event) data structure:

```
# Each patient contributes multiple rows, one per interval
# during which the covariate value is constant
trial_long <- expand_to_intervals(trial, biomarker_data)

fit_tvc <- coxph(Surv(start, stop, event) ~
                   treatment + biomarker_current,
                 data = trial_long)
```

The (start, stop, event) form is also the right structure for clone-censor-weight analyses (Ch 4) and for handling left truncation (when patients enter the risk set after time zero).

Check your understanding: when stratified beats covariate

Question. A trial Cox model includes 'stage' (I, II, III, IV) as a covariate. The PH test on 'stage' is significant — the hazard ratios across stages are not constant over time. Should you include stage as a covariate, as a stratum, or both?

Answer.

Stratify on stage. Including stage as a covariate when its hazard ratio is non-proportional gives a biased estimate of the treatment effect (the model mis-specifies the baseline hazard). Stratifying lets each stage

have its own baseline hazard, and the treatment effect is estimated within strata. The trade is that you lose the ability to estimate the 'effect of stage' from the model — but you would not have wanted that estimate anyway under non- proportional hazards (it is not a single number). Use stage's effect on survival from a separate KM-by-stage analysis or absolute risks.

7.6. Competing risks

A competing risk is an event that prevents the event of interest from occurring. In oncology, death from non-cancer causes competes with death from cancer or disease progression.

The naive Cox analysis treats competing risks as censoring, which assumes patients censored due to death from non-cancer causes have the same future risk of cancer death as those who continue at risk — biologically false. Two correct approaches:

Cause-specific hazards estimate the hazard of each cause separately:

$$\lambda_k(t \mid X) = \lambda_{0k}(t) \exp(X^T \beta_k)$$

Useful when the question is 'what affects the rate of this specific cause of failure'. Estimated by Cox with each cause treated as the event and all others censored:

```
# cause-specific for cancer death
fit_cs_cancer <- coxph(Surv(time, status == "cancer") ~
                         treatment + age,
                       data = trial)

# cause-specific for non-cancer death
fit_cs_non <- coxph(Surv(time, status == "non_cancer") ~
                      treatment + age,
                    data = trial)
```

Fine-Gray subdistribution hazard estimates the 'subdistribution hazard' for one cause, which models the cumulative incidence directly:

$$\bar{\lambda}_k(t \mid X) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t, K = k \mid T \ge t \cup (T < t \cap K \ne k))}{\Delta t}$$

Useful when the question is 'what is the cumulative incidence of this cause' (and you want the model to respect the competing risks).

```
library(cmprsk)

fit_fg <- crr(ftime = trial$time,
              fstatus = trial$status,
              cov1 = trial[, c("treatment", "age")],
              failcode = 1, cencode = 0)
summary(fit_fg)
```

The choice between cause-specific and Fine-Gray depends on the question. Aetiology questions (what affects the rate of this cause) favour cause-specific. Prediction and risk-stratification questions (what is the cumulative incidence) favour Fine-Gray. Modern applied work often reports both (Austin et al., 2016).

The `tidycmprsk` package provides a tidy interface to both.

7.7. Restricted mean survival time

When proportional hazards fails or the audience needs an absolute summary, RMST is the modern alternative (Uno et al., 2014):

$$\text{RMST}(t^*) = E[\min(T, t^*)] = \int_0^{t^*} S(u)\, du.$$

The expected survival time over the follow-up window $[0, t^*]$. Has units of time (e.g., months).

The RMST difference between treatment and control is the difference in expected survival time over the window — interpretable directly as 'patients on treatment lived an average of X months longer over Y months of follow-up'.

```
library(survRM2)

rmst_result <- rmst2(time = trial$time,
                     status = trial$event,
                     arm = trial$treatment,
                     tau = 24)  # 24-month window
print(rmst_result)
#> Restricted Mean Survival Time
#> tau = 24
#> Group 1 RMST: 18.4 (SE 0.8)
#> Group 0 RMST: 16.1 (SE 0.9)
#> Difference  : 2.3 (95% CI 0.7-3.9, p = 0.005)
```

The interpretation: over 24 months, patients on treatment lived an average of 2.3 months longer than control (95% CI 0.7-3.9 months). Compare to the HR of 0.7 — the RMST difference grounds the effect in absolute time, which is what clinicians and patients can act on.

For RMST as a primary summary: `nph::nphHR()` provides similar functionality. The choice of t^* should be specified in the protocol; the maximum observed follow-up time is a common choice but artificial.

7.8. Recurrent events

Some events recur within patients (hospitalisations, infections, exacerbations). Treating only the first event ignores most of the information. Approaches:

- **Anderson-Gill model**: extends Cox to recurrent events using the counting-process framework. Treats events as conditionally independent given covariates.
- **Frailty models** (`coxme`, `frailtypack`): random effects for unobserved patient-level susceptibility.
- **Marginal models** (`survival::coxph` with `cluster`): population-average effects with robust SE.

The `survival` package's vignette on time-dependent covariates and recurrent events is the practical reference.

7.9. Immortal-time bias and the target-trial framework

Immortal-time bias is the most common error in observational survival analysis. Example: a study compares 'patients on drug X for at least 6 months' to 'patients not on drug X'. The drug-X group is guaranteed to have survived 6 months; the comparison group is not. The 'time at risk' is asymmetric, and the apparent benefit of drug X is exaggerated.

The fix: use the (start, stop, event) structure to treat drug-X exposure as a time-varying covariate that turns on when the patient initiates the drug. Or use the target-trial framework (Ch 2) to define exposure groups at time zero.

Other survival-specific traps:

- **Conditioning on the future**: defining a baseline covariate using information observed after time zero. E.g., 'patients who eventually reach disease milestone X' — by conditioning on reaching the milestone, you condition on a function of the outcome.
- **Differential follow-up**: when treatment groups have different follow-up durations (e.g., one arm enrolled later), the 'censoring' patterns differ and bias may result.

These are deep traps; the careful biostatistician recognises them before fitting the model.

7.10. Worked example: a survival analysis with all the pieces

A trial of treatment X vs. standard care in metastatic cancer. Outcome: overall survival. Competing risk: death from non-cancer causes. Median follow-up: 24 months.

```
library(survival)
library(survminer)
library(tidycmprsk)
library(survRM2)

trial <- read_csv("data/cancer-trial.csv")

# 1. KM curves with risk table
ggsurvplot(survfit(Surv(time, event) ~ treatment,
                   data = trial),
           data = trial,
           risk.table = TRUE,
           conf.int = TRUE,
           pval = TRUE)

# 2. Cox PH model
fit_cox <- coxph(Surv(time, event) ~ treatment + age +
                   sex + ecog,
                 data = trial)
summary(fit_cox)

# 3. Check PH assumption
cox.zph(fit_cox)
# Suppose the test on 'stage' is significant —
# stratify
fit_cox2 <- coxph(Surv(time, event) ~ treatment + age +
                    sex + ecog + strata(stage),
                  data = trial)
cox.zph(fit_cox2)  # all p > 0.05

# 4. Competing-risks analysis
fit_cs <- cuminc(Surv(time, status) ~ treatment,
                 data = trial)
plot(fit_cs)

# 5. RMST as a sensitivity summary
rmst2(time = trial$time, status = trial$event,
```

```
       arm = trial$treatment, tau = 24)
```

The reported analysis includes:

- KM curves and the log-rank p-value as the primary visual.
- Cox HR (with appropriate stratification) for the primary effect estimate.
- RMST difference at 24 months as a clinically interpretable summary.
- Competing-risks cumulative incidence to address the elderly population (where non-cancer death is substantial).

The methods section names each: 'we report the Cox hazard ratio, with stratification on stage to address non-proportional hazards. The treatment effect is also summarised as a 24-month RMST difference and as cumulative incidence functions to address competing risks from non-cancer death.'

7.11. Collaborating with an LLM on applied survival analysis

Three patterns.

Prompt 1: 'Build the survival model for this trial.' Provide the data structure and the protocol.

What to watch for. The LLM produces a clean Cox formulation. It often defaults to including all covariates without stratification; if PH fails on a covariate, the model is biased. Push for the PH check and stratification or alternative summary.

Verification. Run `cox.zph()` on the proposed model; inspect the result.

Prompt 2: 'Should we use Fine-Gray or cause-specific for this competing-risks question?' Provide the question and the data.

What to watch for. The LLM gives a competent distinction. It tends to default to one or the other without anchoring to the substantive question. Push for the link between question and method.

Verification. Map the choice back to the question. Aetiology favours cause-specific; risk prediction favours Fine-Gray.

Prompt 3: 'Is this analysis vulnerable to immortal-time bias?' Provide the analysis description.

What to watch for. The LLM correctly identifies classic immortal-time-bias patterns ('on-treatment for X days as baseline covariate'). It tends to miss subtler versions ('survived to receive a particular test result'). Provide enough detail for the LLM to judge.

Verification. The clearest test: is the exposure defined using information observed after time zero? If yes, you have the bias.

The meta-pattern: LLMs are good at the syntactic mechanics (writing the Cox formula, suggesting diagnostics) and weak at the substantive judgement (whether PH holds in your data, which framework matches the question, whether immortal time is present). Use them for code; bring the substantive reasoning yourself.

7.12. Principle in use

Three habits.

1. **Always check proportional hazards.** `cox.zph()` is one line; report the result.
2. **Address competing risks deliberately.** Cause-specific or Fine-Gray, chosen by question. Both reported when the audience benefits.
3. **Pair the HR with an absolute summary.** The HR is fine for the primary; RMST or milestone survival makes the magnitude interpretable.

7.13. Exercises

1. For a published trial in your area, compute the RMST difference at the trial's primary follow-up point. Compare to the reported HR. Reflect on which is the better summary for a clinical audience.

2. Take a competing-risks dataset (e.g., the `bmt` data in `cmprsk`). Compute cumulative incidence functions for both events. Compare to the naive 1-KM estimate; quantify the overestimation.
3. Run `cox.zph()` on a Cox model from your work. For any covariate with a significant non-proportionality, fit a stratified version and compare the treatment HR.
4. Construct a hypothetical immortal-time bias situation. Estimate the magnitude of the bias by simulation. Document.
5. Implement an analysis with time-varying covariates using the (start, stop, event) structure. Verify the result against a single-baseline-covariate model on a dataset without changes.

7.14. Further reading

- Therneau & Grambsch (2000), *Modeling Survival Data: Extending the Cox Model*. The reference for the applied survival material in this chapter.
- Kleinbaum & Klein (2012), *Survival Analysis: A Self-Learning Text*. The applied textbook.
- Austin et al. (2016), 'Practical recommendations for reporting Fine-Gray model analyses for competing risk data'. The applied reference.
- Uno et al. (2014), 'Moving beyond the hazard ratio in quantifying the between-group difference'. The RMST reference.
- The `survival`, `cmprsk`, `tidycmprsk`, `survminer`, and `survRM2` package documentation are the practical references.

Part IV.

Clinical Trials

8. Clinical Trial Design

8.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish trial phases (I, II, III, IV) and identify the design questions specific to each.
- Choose between simple, stratified, block, and covariate-adaptive randomisation, and recognise the consequences of each for analysis.
- Conduct a sample-size calculation for a two-arm trial with a continuous, binary, or time-to-event endpoint, including non-inferiority designs.
- Apply the ICH E9 R1 estimand framework to a proposed trial protocol and identify the primary estimand attributes.
- Recognise when an adaptive, group-sequential, or pragmatic design is the right choice and articulate the trade-offs.

8.2. Orientation

Clinical trial design is where the biostatistician makes some of the most consequential decisions in the project: the primary endpoint, the sample size, the randomisation procedure, the stopping rules, the estimand. Each decision constrains the trial's ability to answer its question; collectively they determine whether the trial succeeds or produces ambiguous data.

This chapter covers design; Chapter 9 covers analysis and reporting. The split mirrors the temporal flow of trial work: design happens before any patient is enrolled and is documented in the protocol; analysis happens after enrolment is complete and is documented in the SAP. Design decisions cannot be undone after the trial starts.

The chapter is organised in three threads. **Design fundamentals**: phases, randomisation, blinding, sample size. **The estimand framework** (continuing from Ch 1): how to specify what the trial estimates. **Modern designs**: adaptive, group-sequential, pragmatic, platform.

8.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) The primary endpoint defines the trial. Everything else — sample size, analysis, stopping rules — flows from it. The biostatistician chooses the primary endpoint by considering three things: clinical relevance (does it inform a decision), measurability (can it be reliably ascertained in this trial), and statistical efficiency (how much information does it contain per patient). A composite primary endpoint is sometimes the right answer, but introduces interpretation complexity; a continuous endpoint is sometimes the right answer, but may not match the regulatory question.

(Judgement 2.) Sample size is a contract with the data. The sample size is computed from the primary analysis's effect size, variability, type I error, and power. The biostatistician's role is to defend each input, not to engineer the calculation to produce a feasible n. An optimistic effect-size assumption produces a trial powered to detect an effect that does not exist; a pessimistic one produces a trial too large to enrol.

(Judgement 3.) Estimands precede randomisation. The estimand framework (Ch 1) applies in trials with particular force because intercurrent events (treatment discontinuation, switching, death) are known in advance to occur. The biostatistician specifies the primary estimand and the intercurrent-event strategy in the protocol, before any data is collected. ITT is a treatment-policy strategy; it is one estimand among several, not the default for every trial.

These judgements are what distinguish a trial that informs a regulatory or clinical decision from a trial that produces ambiguous data.

8.4. Phases of clinical development

Phase I: first-in-human, dose-finding. Small (often 20-80 patients), often healthy volunteers (oncology exception). Primary endpoint typically safety, pharmacokinetics. Designs: 3+3, model-based (CRM, EWOC, BOIN), Bayesian-adaptive.

Phase II: efficacy signal in patient population. Single-arm or small randomised, 50-200 patients. Primary endpoint often a surrogate (response rate, biomarker change, PFS) rather than a hard endpoint. Designs: Simon's two-stage, Bayesian-adaptive.

Phase III: definitive efficacy. Large randomised (hundreds to thousands), powered for hard clinical endpoints. The phase that supports regulatory approval. Standard designs: parallel-group, two-arm; sometimes multi-arm or factorial.

Phase IV: post-approval surveillance and real-world effectiveness. Pragmatic designs, often embedded in routine care. Outcomes include long-term safety, real-world effectiveness, comparative effectiveness.

The chapter focuses on Phase III conventions, since those are most fully developed and most commonly encountered in MS-level biostatistical work. Phase I designs require a separate methodology (the 'A Practical Guide to Biostatistics' literature) beyond the scope here.

8.5. Randomisation

The point of randomisation: balance unmeasured confounders in expectation, eliminate selection bias by the investigator, create a reference distribution for the test statistic.

Simple randomisation: each patient assigned independently with probability 0.5 (for a 1:1 trial). Easy; can produce imbalance in small trials.

Block randomisation: assign in blocks of size 2, 4, or 6 (or vary the block size). Guarantees balance at the end of each block; reduces imbalance from chance. Standard for most trials.

Stratified randomisation: block randomisation within strata defined by one or more baseline variables (e.g., centre, disease stage). Guarantees balance within strata. Use when the strata are strongly prognostic and the trial is small enough that imbalance within a stratum could matter.

Covariate-adaptive randomisation (e.g., minimisation): assigns each new patient based on the current imbalance across multiple baseline variables. Achieves better balance than stratified randomisation but introduces a degree of predictability that some regulatory bodies discourage. The literature is mixed (Taves, 2010); check current FDA / EMA guidance.

Implementation: most trials use a centralised randomisation system (web or IVRS) that handles the mechanics. The biostatistician specifies the procedure in the protocol; an unblinded statistician (separate from the analysis team) sometimes implements and monitors.

The analysis must respect the randomisation procedure. A stratified randomisation requires stratified analysis; ignoring the stratification inflates the type I error.

8.6. Blinding

Blinding addresses bias from differential treatment or assessment of outcomes:

- **Single-blind**: the patient does not know their assignment.
- **Double-blind**: neither patient nor investigator knows.
- **Triple-blind**: patient, investigator, and analysis team are blinded.

Blinding is rarely fully achievable for invasive treatments (surgery, devices) and is sometimes infeasible for behavioural interventions. When blinding fails, the analysis should examine investigator effects and consider sensitivity analyses.

The unblinding procedure is specified in advance: typically after the database lock, after the analysis of the primary endpoint by the unblinded statistician, or after the last patient's last visit.

8.7. Sample size

The two-arm continuous-outcome calculation:

$$n_{\text{per arm}} = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

where σ is the within-arm SD, δ is the target effect size (mean difference), and z are standard normal quantiles. For two-sided 5% type I error and 80% power, $z_{1-\alpha/2} + z_{1-\beta} = 1.96 + 0.84 = 2.80$, and:

$$n_{\text{per arm}} \approx 2 \cdot 7.85 \cdot (\sigma/\delta)^2.$$

Concrete: with $\sigma = 10$ mmHg and $\delta = 5$ mmHg, $n \approx 63$ per arm.

The binary-outcome calculation:

$$n_{\text{per arm}} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2}.$$

The time-to-event calculation depends on the expected number of events, not the number of patients:

$$\text{events} = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{\log(\text{HR})^2}.$$

The R packages `pwr`, `clinfun`, `gsDesign`, and `pwrss` implement these formulas plus more elaborate ones. The `Hmisc::cpower()` and the dedicated trial- design package `rpact` are alternatives.

```
library(pwr)

# two-sample t-test
pwr.t.test(d = 5/10, sig.level = 0.05, power = 0.8,
           type = "two.sample")
#> n = 63.8 per arm

# proportions
pwr.2p.test(h = ES.h(0.30, 0.20),
            sig.level = 0.05, power = 0.8)
#> n = 313.5 per arm
```

The biostatistician's task: defend each input. Where does σ come from (a prior trial? the literature? expert opinion)? Where does δ come from (clinically meaningful difference? smallest worth detecting?)? Why this α (one-sided vs. two-sided?), this power (80%? 90%?). The protocol should answer each.

8.8. Non-inferiority and equivalence

A standard trial tests whether treatment X is **better** than control at a pre-specified α. Sometimes the question is whether X is **not worse** by more than a pre-specified margin. The non-inferiority margin Δ is the largest acceptable difference in the wrong direction.

The hypothesis:

$$H_0 : \mu_X - \mu_C \leq -\Delta \text{ vs. } H_1 : \mu_X - \mu_C > -\Delta$$

(for outcomes where higher is better).

Sample size for non-inferiority is generally larger than for superiority because the margin is smaller than the typical effect size. The power calculation uses the formula above with δ replaced by the effect on the non-inferiority margin (zero if the new treatment is assumed identical, the actual expected effect minus the margin if some superiority is expected).

The non-inferiority margin is a substantive choice: the largest difference that would still be clinically acceptable. It should be defended on clinical grounds and is typically a fraction (often 50%) of the historical effect size of the active control over placebo, to preserve at least half of the historical benefit. The biostatistician proposes; the clinical team decides.

Check your understanding: when non-inferiority is the right design

Question. A new oral anticoagulant is proposed as a more convenient alternative to warfarin (which requires INR monitoring). The new drug is expected to have similar efficacy. Should the trial be designed as superiority, non-inferiority, or equivalence?
Answer.
Non-inferiority is the right design. The clinical question is 'is the

new drug at least as good as warfarin in terms of efficacy, given its convenience advantage?' Superiority would require demonstrating that the new drug is better, which is unlikely (and not necessary for the clinical use case). Equivalence would require demonstrating the new drug is neither better nor worse, which is rarely needed; for a 'similar efficacy, better convenience' positioning, the asymmetric non-inferiority test is sufficient and better-powered. The non-inferiority margin would be a fraction (typically 50%) of the warfarin-vs-no-anticoagulant effect on the primary endpoint (stroke), to preserve at least half of warfarin's historical benefit.

8.9. Estimand framework for trials

The ICH E9 R1 estimand framework (Ch 1) applies to trials with particular force. The five attributes (population, treatment, comparator, outcome, summary) plus the intercurrent-event strategy must be specified in the protocol.

Common intercurrent events in trials:

- Treatment discontinuation (patients stop taking the treatment)
- Treatment switching (patients switch to the other arm or to a third treatment)
- Death (when not the primary outcome)
- Use of rescue medication
- Loss to follow-up

For each, the protocol specifies one of the five strategies (treatment policy, composite, hypothetical, principal stratum, while-on-treatment).

The default for many regulatory trials is treatment-policy for treatment discontinuation (estimating the effect of being assigned to a treatment strategy regardless of adherence). This is ITT. Per-protocol or hypothetical analyses are secondary or sensitivity. The protocol must name the primary estimand explicitly.

For a non-inferiority trial, ITT may be the wrong primary because it biases toward no difference (any 'noise' from non-adherence makes the

new treatment look more like the control). Per-protocol or hypothetical strategies are sometimes the primary in non-inferiority trials.

8.10. Adaptive and group-sequential designs

A **group-sequential** design analyses the data at pre-specified interim time points and may stop the trial early for efficacy or futility. The standard spending functions (O'Brien-Fleming, Pocock, Lan-DeMets) control the overall type I error across multiple looks.

The `gsDesign` R package implements these:

```
library(gsDesign)

design <- gsDesign(k = 4,                # 4 looks
                   test.type = 4,         # 2-sided
                   alpha = 0.025, beta = 0.1,
                   sfu = sfLDOF,         # O'Brien-Fleming spending
                   sfl = sfLDPocock)     # Pocock for futility
summary(design)
```

The output: the cumulative information times at each look, the boundary values (in standard units), expected sample size under the null and alternative.

An **adaptive** design changes some aspect of the trial based on accumulated data: sample size, allocation ratio, the primary endpoint, the patient population. Adaptive designs are increasingly common in oncology (basket and umbrella trials, platform trials) but require careful pre-specification to maintain the overall type I error.

The FDA's guidance on adaptive designs (U.S. Food and Drug Administration, 2019) is the authoritative reference; the methodological literature is rich and evolving.

8.11. Pragmatic and platform trials

Pragmatic trials answer real-world effectiveness questions: enroll broad populations, deliver treatment in routine care, measure outcomes from existing data sources. Trade efficiency for generalisability. The PRECIS-2 framework (Loudon et al., 2015) is the design tool.

Platform trials test multiple interventions simultaneously against a shared control, with the ability to drop arms for futility and add new arms mid-trial. STAMPEDE in prostate cancer and the RECOVERY trial in COVID-19 are canonical examples. Platform trials require sophisticated statistical methods (master protocols, Bayesian decision rules) and substantial operational infrastructure.

For an MS-level biostatistician, knowing the names and the structural ideas of pragmatic and platform trials is enough; implementing one requires specialist methodological involvement.

8.12. Worked example: designing a Phase III trial

A new drug for moderate-to-severe COPD is ready for Phase III. Active control is salmeterol/fluticasone. Primary endpoint: change in FEV1 from baseline to week 24.

Step 1. Estimand. - Population: adults 40+ with moderate-to-severe COPD. - Treatment: new drug (specific dose). - Comparator: salmeterol/fluticasone. - Outcome: change in trough FEV1 from baseline to week 24 (mL). - Population summary: mean difference between arms. - Intercurrent events: - Treatment discontinuation: treatment policy (analyse as randomised). - Use of rescue medication: while-on-treatment (analyses based on FEV1 measurements during on-treatment period).

Step 2. Sample size. Prior trials show within-arm SD of 200 mL and a clinically meaningful difference of 50 mL. With $\alpha = 0.05$ two-sided and 90% power:

$$n_{\text{per arm}} = \frac{2 \cdot 200^2 \cdot (1.96 + 1.28)^2}{50^2} \approx 339.$$

Add 15% for dropout: 390 per arm. Total enrolment: 780.

Step 3. Randomisation. Stratified by baseline disease severity (moderate vs. severe) and current ICS use (yes/no). Block size 4. Centralised IVRS.

Step 4. Blinding. Double-blind, double-dummy. Both arms receive an identical-looking inhaler.

Step 5. Interim analysis. Single interim at 50% of events, with O'Brien-Fleming boundary for early efficacy stopping. Sample-size re-estimation based on the observed within-arm SD (without unblinding the treatment effect).

Step 6. Analysis. Mixed model for repeated measures (MMRM), with fixed effects for treatment, visit, treatment-by-visit, baseline FEV1, baseline severity, ICS use; unstructured covariance for repeated visits within patient. Handle missing data under MAR (the MMRM assumption); sensitivity analyses for MNAR (Ch 10).

The protocol now contains everything the trial needs. Six months in, when the DSMB asks 'what was the sample-size justification?', the protocol is the answer.

8.13. Collaborating with an LLM on clinical trial design

Three patterns.

Prompt 1: 'Compute the sample size for this trial.' Provide effect size, variability, alpha, power.

What to watch for. The LLM produces a calculation using `pwr` or `gsDesign`. It often defaults to two-sided α when one-sided is appropriate (or vice versa) and over-simplifies the assumptions. Push for explicit defence of each input.

Verification. Recompute by formula or with a second package. The arithmetic is elementary and worth double-checking.

Prompt 2: 'Write the estimand for this trial.' Provide the trial protocol summary.

What to watch for. The LLM produces all five attributes. It commonly under-specifies the intercurrent-event strategies. Push for explicit naming of each strategy.

Verification. Read against ICH E9 R1.

Prompt 3: 'Should this trial use a group-sequential design?' Provide the trial details.

What to watch for. The LLM gives a competent discussion. It tends to recommend group-sequential designs more often than warranted (the operational cost is real). Push for the trade-offs.

Verification. The decision should consider the cost of an interim (operational, regulatory submissions, DSMB), not just the statistical efficiency.

The meta-pattern: LLMs are good at the syntactic mechanics (writing the formula, drafting the protocol section) and weak at the substantive choices (which estimand, which interim strategy, which margin in non-inferiority). Use them for code and drafts; bring substantive judgement yourself.

8.14. Principle in use

Three habits.

1. **Estimand before sample size.** The sample size is computed for a specific estimand and a specific estimator. Specifying the estimand first prevents the calculation from being detached from the primary analysis.

2. **Defend each sample-size input.** Effect size, variability, alpha, power, dropout rate. Each number in the calculation has a source; the protocol cites it.

3. **Document intercurrent-event strategies in the protocol.** ITT is one strategy among several; the choice should be explicit.

8.15. Exercises

1. For a hypothetical trial in your area, write the primary estimand using all six ICH E9 R1 attributes. Identify the intercurrent-event strategies for at least two intercurrent events.
2. Compute the sample size for a two-arm continuous-outcome trial with $\sigma = 15$, $\delta = 5$, $\alpha = 0.05$ two-sided, power $= 80\%$. Verify by formula.
3. For a published Phase III trial, find the sample- size justification in the protocol or supplement. Identify the inputs (effect size, variability, etc.) and assess whether each is defended.
4. Design a non-inferiority trial. Choose a margin Δ and defend it on clinical grounds. Compute the sample size required.
5. Use `gsDesign` to design a 4-look group-sequential trial with O'Brien-Fleming efficacy boundary. Report the boundaries at each look and the expected sample size under the null.

8.16. Further reading

- International Council for Harmonisation (2019), *ICH E9(R1) Addendum on Estimands and Sensitivity Analysis*. The regulatory text.
- Piantadosi (2017), *Clinical Trials: A Methodologic Perspective* (3rd edition). The reference textbook.
- Friedman et al. (2015), *Fundamentals of Clinical Trials* (5th edition). The classical applied reference.
- Jennison & Turnbull (2000), *Group Sequential Methods with Applications to Clinical Trials*. The reference for group-sequential design.
- The `gsDesign`, `rpact`, `pwr`, and `clinfun` package documentation.

9. Clinical Trial Analysis and Reporting

9.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish ITT, modified-ITT, per-protocol, and as-treated analysis populations and choose the primary analysis that matches the trial's primary estimand.
- Apply standard adjustment strategies (stratified analysis, ANCOVA, MMRM) and recognise when each is appropriate.
- Implement multiplicity control for multiple endpoints, multiple comparisons, and interim analyses (Bonferroni, Hochberg, gatekeeping, spending functions).
- Draft a CONSORT-compliant trial report including the flow diagram, baseline-characteristics table, primary analysis, and pre-specified subgroup analyses.
- Conduct ICH E9 R1-style sensitivity analyses for the primary estimand.

9.2. Orientation

Chapter 8 covered the design decisions made before enrolment. This chapter covers the analysis decisions made after database lock, and the reporting conventions the trial must satisfy. The boundary is not perfectly clean — the SAP is finalised before unblinding — but the methodological focus is.

The chapter is organised in three threads. **Analysis populations**: ITT, mITT, PP, AT, and the connection to the primary estimand. **Adjustment**

and multiplicity: stratified analysis, ANCOVA, MMRM, multiplicity-controlled testing. **Reporting**: the CONSORT flow diagram, the baseline-characteristics table, the primary-analysis presentation, sensitivity analyses. The chapter closes with the regulatory expectations as of 2026.

9.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) The primary analysis matches the primary estimand. The protocol specifies the primary estimand; the SAP specifies the primary analysis. The analysis must estimate the estimand. ITT for treatment-policy estimands; PP or hypothetical for adherence-conditional estimands. The biostatistician identifies any discrepancy between primary estimand and primary analysis before unblinding and corrects it.

(Judgement 2.) Pre-specification is a discipline, not a formality. Every analysis decision in the SAP constrains the analysis team. Decisions made after unblinding (or after seeing the data) introduce flexibility that biases the result toward whatever hypothesis the analyst preferred. The biostatistician maintains the SAP as a contract: deviations are documented and justified, not silent.

(Judgement 3.) Sensitivity analyses are part of the result. The primary analysis estimates the primary estimand under the primary assumptions. Sensitivity analyses estimate the same estimand under perturbed assumptions (different missing-data mechanism, different intercurrent-event strategy, different covariates). The biostatistician designs the sensitivity analyses with the primary, not after the primary result is in.

These judgements are what distinguish a defensible trial report from one that produces a number with inadequate context.

9.4. Analysis populations

Four standard populations:

Intention-to-treat (ITT): every randomised patient analysed in the arm to which they were assigned, regardless of treatment received. Estimates the treatment-policy estimand. The most conservative choice for superiority trials (deviations from the assigned treatment dilute toward the null); typically the primary for regulatory superiority trials.

Modified ITT (mITT): ITT excluding specific subsets (e.g., patients who never received any dose, patients with no post-baseline measurement). The exclusions must be pre-specified and justified; otherwise mITT becomes a vehicle for selecting the best-looking analysis.

Per-protocol (PP): only patients who received the assigned treatment as specified, with adequate adherence and no major protocol deviations. Estimates closer to the hypothetical estimand. Sometimes the primary in non-inferiority trials.

As-treated (AT): patients analysed in the arm of the treatment they actually received. Useful for safety analyses; rarely appropriate as the primary efficacy population.

The choice of population is part of the protocol; it is not made post-hoc. Every analysis defines its population precisely.

9.5. Adjustment strategies

Stratified analysis when randomisation was stratified. The stratification variables enter the analysis as fixed effects (or as a stratified test statistic for non-parametric methods). Failing to stratify when the design did inflates type I error and reduces power.

ANCOVA (Analysis of Covariance) for continuous endpoints with a pre-specified baseline covariate (the baseline value of the endpoint, in particular). Adjusting for baseline removes a known source of variability, narrows the CI for the treatment effect, and is the default for change-from-baseline analyses.

```
fit_ancova <- lm(week24 ~ treatment + baseline + stratum,
                 data = trial)
summary(fit_ancova)
```

ANCOVA on the change score (`week24 - baseline ~ treatment`) gives a different result from ANCOVA on the raw endpoint. The latter is preferred (regression to the mean is properly handled).

MMRM (Mixed Model for Repeated Measures) for longitudinal continuous endpoints. The standard for trials with repeated measurements:

```
library(nlme)

fit_mmrm <- gls(value ~ visit * treatment + baseline +
                  stratum,
                data = trial_long,
                correlation = corSymm(form = ~ visit_num | id),
                weights = varIdent(form = ~ 1 | visit))
```

The `corSymm` allows arbitrary correlation across visits; `varIdent` allows different variance at each visit. This unstructured covariance is the standard choice and is robust to mis-specification.

The `mmrm` package provides a more user-friendly interface specific to clinical trials.

For binary or categorical endpoints, the standard adjustments are stratified Cochran-Mantel-Haenszel or logistic regression with covariates. For time-to-event, stratified Cox regression.

The pre-specified covariates and stratification factors should be listed in the SAP. Adding covariates post-hoc is a deviation that requires justification and is typically reported as a sensitivity rather than the primary.

9.6. Multiplicity

Multiple comparisons inflate type I error if not controlled. Four contexts:

Multiple primary endpoints. Two or more endpoints that must each be statistically significant for the trial to succeed (the **co-primary** structure) — no multiplicity adjustment needed for the conjunction because the joint type I error is already controlled. Or one endpoint that 'wins' if any is

significant (the **multiple-primary** structure) — multiplicity adjustment required.

Multiple secondary endpoints. Hierarchical testing (test endpoint 2 only if endpoint 1 is significant) controls familywise error without explicit adjustment. Bonferroni or Hochberg adjustment if all are tested simultaneously.

Multiple comparisons across treatment arms. In a multi-arm trial, comparing each new treatment to control requires adjustment (Dunnett's procedure is standard).

Multiple looks (interim analyses). Group-sequential boundaries (Ch 8) control type I error across interim analyses.

The SAP specifies the multiplicity strategy explicitly. Common patterns:

- **Bonferroni**: simple, conservative. Divide α by the number of tests. α/k.
- **Holm**: step-down. Test the most significant first at α/k, then $\alpha/(k-1)$, etc. Less conservative than Bonferroni; uniformly more powerful.
- **Hochberg**: step-up. Test the least significant first at α, then $\alpha/2$, etc. More powerful than Holm under PRDS.
- **Hierarchical (gatekeeping)**: test endpoints in pre-specified order; subsequent tests only if all prior pass. No formal adjustment.

The `multcomp` and `gMCP` packages implement these in R.

9.7. CONSORT and reporting standards

The CONSORT 2010 guidelines (Schulz et al., 2010) are the reference for trial reporting. The most-cited elements:

The CONSORT flow diagram: shows enrolment, randomisation, allocation, follow-up, and analysis counts at each stage. Required by most journals.

The `consort` R package generates the diagram programmatically:

```
library(consort)
g <- consort_plot(...)
```

Table 1 (baseline characteristics by treatment arm) demonstrates the success of randomisation. It is a description, not a hypothesis test; the convention is to NOT report p-values for differences in baseline variables in randomised trials (Knol et al., 2012). The `tableone` and `gtsummary` packages produce publication-ready tables.

```
library(gtsummary)
trial |>
  select(treatment, age, sex, bmi, severity) |>
  tbl_summary(by = treatment) |>
  add_overall()
```

Primary analysis presentation: the point estimate, its 95% CI, the p-value (where applicable), the analysis method, the population, the intercurrent-event strategy. A complete sentence:

> 'In the ITT population, treatment with X reduced mean change in HbA1c at week 24 by 0.42% (95% CI 0.31-0.53%, $p < 0.001$) compared with placebo, using MMRM with treatment, visit, treatment-by-visit, and baseline HbA1c as fixed effects, with unstructured covariance.'

Subgroup analyses: pre-specified subgroups (age, sex, baseline severity) with forest plots. Reported as point estimates and CIs, not as subgroup p-values. The `forestplot` package draws these.

Sensitivity analyses: alternative intercurrent-event strategies, alternative missing- data assumptions (Ch 10), alternative populations. Each reported with its rationale.

9.8. ICH E9 R1 sensitivity analyses

The estimand framework (Chs 1, 8) demands sensitivity analyses. Specifically:

- **Sensitivity to the intercurrent-event strategy**: if treatment-policy is primary, also report hypothetical (treating discontinuation as if it did not happen). The discrepancy is informative about adherence.
- **Sensitivity to missing-data assumption**: primary under MAR (the MMRM assumption); sensitivity under MNAR (Ch 10). Pattern-mixture or selection-model-based.
- **Sensitivity to model specification**: primary with the planned covariate set; sensitivity with a wider or narrower set.

The sensitivity analyses are pre-specified in the SAP. The reporting separates the primary result from the sensitivities; the latter contextualise the former.

Check your understanding: when ITT is not the right primary

Question. A non-inferiority trial of a new anticoagulant vs. warfarin for preventing stroke. The non-inferiority margin is 1.38 on the HR scale (the new drug must not be more than 38% worse). Should the primary analysis be ITT or per-protocol?
Answer.
For non-inferiority, **per-protocol** (or hypothetical) is often preferred as the primary, with ITT as sensitivity. The reason: ITT in non-inferiority biases toward the null (no difference), which makes it *easier* to declare non-inferiority. A per-protocol analysis, where adherence is enforced, is more conservative for the non-inferiority claim. The typical regulatory pattern is to require both ITT and PP analyses, with consistency between them required to declare non-inferiority. The protocol must specify the choice and defend it.

9.9. Bayesian analyses in trials

Modern regulatory trials increasingly include Bayesian analyses, either as primary (in some device trials and adaptive trials) or as sensitivity. The framework:

- A pre-specified prior (often weakly informative or reference).
- Posterior distribution of the treatment effect.

- The 'success criterion' is a posterior probability threshold (e.g., $\Pr(\text{effect} > 0) > 0.975$).

The FDA's guidance on Bayesian methods (U.S. Food and Drug Administration, 2010) is the regulatory reference. The methodology is well-developed; the operational challenge is pre-specification of the prior. The prior should be defended with reference to historical data or via robustness checks across priors.

```
library(brms)

fit_bayes <- brm(value ~ visit * treatment + baseline,
                 data = trial,
                 prior = c(prior(normal(0, 1), class = "b")),
                 chains = 4, cores = 4)
summary(fit_bayes)
```

For most MS-level applied work, frequentist analyses remain the primary; Bayesian methods are sensitivity or specialty. The introductory SCAI volume's Bayesian chapter and the SCAI-advanced volume's MCMC and Modern Bayesian chapters provide the computing foundation.

9.10. Worked example: analysing a Phase III diabetes trial

A trial has randomised 800 patients (400 per arm) to treatment X or placebo for 24 weeks. Primary endpoint: change in HbA1c from baseline to week 24.

```
library(tidyverse)
library(mmrm)
library(gtsummary)

trial <- read_csv("data/diabetes-trial.csv")

# 1. CONSORT flow numbers
```

```
trial |>
  count(stage = "Randomised", treatment)
trial |>
  filter(received_treatment == 1) |>
  count(stage = "Received treatment", treatment)
# (etc.)

# 2. Table 1 by arm
tbl_summary(trial |> filter(visit == 0) |>
              select(treatment, age, sex, bmi,
                     hba1c, sbp, dbp),
            by = treatment) |>
  add_overall()

# 3. Primary analysis (MMRM, ITT)
fit_primary <- mmrm(
  formula = hba1c ~ treatment + visit +
              treatment:visit + baseline_hba1c +
              us(visit | id),
  data = trial_long
)
summary(fit_primary)

# extract the contrast at week 24
emmeans::emmeans(fit_primary, ~ treatment | visit,
                 at = list(visit = 24))
emmeans::contrast(...)

# 4. Sensitivity: per-protocol
fit_pp <- mmrm(formula = ..., data = trial_pp)

# 5. Sensitivity: jump-to-reference for missing data
# (Ch 10 develops the methodology)
fit_jr <- ...

# 6. Pre-specified subgroup forest plot
subgroups <- c("age_group", "sex", "baseline_hba1c_q",
```

```
               "duration_diabetes")
forest_data <- subgroups |> map_dfr(...)
forestplot(forest_data)

# 7. Safety summary (different population: AT)
tbl_summary(trial_at |> select(treatment, ae_any,
                               ae_serious, discontinuation),
            by = treatment) |>
  add_p()
```

The methods section reads:

> The primary analysis was performed in the ITT population using mixed model for repeated measures with fixed effects for treatment, visit, treatment-by-visit, baseline HbA1c, and the stratification factor (baseline severity). An unstructured covariance for repeated measurements was specified, and Kenward-Roger degrees of freedom were used. Missing data were handled under missing-at-random assumptions implicit in the MMRM. Sensitivity analyses included a per-protocol analysis and a jump-to-reference imputation under MNAR. Pre-specified subgroup analyses were performed; no adjustment was made for multiplicity across subgroups.

The structure is the regulatory standard.

9.11. Collaborating with an LLM on clinical trial analysis

Three patterns.

Prompt 1: 'Draft the SAP analysis section for this trial.' Provide the protocol summary.

What to watch for. The LLM produces a clean draft. It commonly underspecifies the covariate set, the multiplicity strategy, and the sensitivity analyses. Push back on each.

Verification. The SAP is reviewed by senior biostatistician and the regulatory affairs team. LLM drafts accelerate the review; they do not substitute for it.

Prompt 2: 'Implement MMRM for this trial.' Provide the data structure and the SAP.

What to watch for. The LLM produces working code using `mmrm` or `nlme::gls`. It often defaults to a simpler covariance structure than unstructured. Verify against the SAP.

Verification. The output should match the pre-specified analysis exactly. Any deviation is a deviation from the SAP and must be documented.

Prompt 3: 'Generate the CONSORT flow diagram for this trial.' Provide the screening, randomisation, follow-up, and analysis counts.

What to watch for. The LLM produces working code using the `consort` package. The numbers must match the SAP and the actual data; verify.

Verification. The flow diagram is double-checked against the SAP. Discrepancies are a serious problem that must be resolved before publication.

The meta-pattern: LLMs are good for the syntactic mechanics (writing the MMRM call, drafting the SAP) and weak at the substantive judgement (whether the analysis matches the estimand, whether the sensitivity is appropriate). Use them for code and drafts; bring substantive judgement yourself.

9.12. Principle in use

Three habits.

1. **Match analysis to estimand.** The primary analysis estimates the primary estimand. ITT is appropriate for treatment-policy; per-protocol or hypothetical for other estimands.
2. **Pre-specify everything.** Every analysis decision is in the SAP before unblinding. Deviations are documented and justified.
3. **Sensitivity analyses are part of the result.** Report them alongside the primary, not as an afterthought.

9.13. Exercises

1. Take a published Phase III trial in your area. Identify the primary estimand, the primary analysis, and the analysis population. Are they consistent?
2. For a hypothetical trial with three primary endpoints, design a multiplicity strategy. Compare Bonferroni, Hochberg, and a hierarchical (gatekeeping) approach in terms of power.
3. Implement MMRM on a longitudinal trial dataset. Compute the contrast at the final visit and compare to a simple ANCOVA on the change score from baseline.
4. Generate a CONSORT flow diagram from a trial dataset. Verify each number against the trial protocol.
5. Conduct a tipping-point sensitivity analysis (Ch 10 develops this) for a trial with 8% missing data on the primary endpoint. Identify the smallest delta that overturns the conclusion.

9.14. Further reading

- International Council for Harmonisation (2019), *ICH E9(R1) Addendum*. The estimand framework as it applies to trial analysis.
- Schulz et al. (2010), 'CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials'. The reporting reference.
- Piantadosi (2017), *Clinical Trials: A Methodologic Perspective*. The reference textbook.
- Committee for Proprietary Medicinal Products (2002), EMA Guideline on Multiplicity Issues in Clinical Trials. Authoritative on multiplicity for European regulatory submissions.
- The `mmrm`, `gtsummary`, `consort`, and `forestplot` packages are the contemporary applied tools.

Part V.

Specialised Methods

10. Missing Data at Depth

10.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish MCAR, MAR, and MNAR mechanisms and recognise each in applied contexts.
- Apply Rubin's framework: multiple imputation, Rubin's rules for combining estimates and variances.
- Implement multiple imputation by chained equations (FCS) with the `mice` package, including predictive-mean-matching for continuous variables and logistic for binary.
- Conduct pattern-mixture and selection-model sensitivity analyses for MNAR.
- Apply the tipping-point analysis for the primary endpoint in a clinical trial.

10.2. Orientation

The practicum volume's missing-data chapter (Ch 17) covered the basics: diagnose the pattern, choose a strategy, document. This chapter develops the methodology at the depth required for trials and publications: Rubin's framework rigorously, multiple imputation done well, sensitivity analyses for MNAR.

The chapter is organised in three threads. **Theory**: MCAR / MAR / MNAR, Rubin's rules. **Practice**: multiple imputation with `mice`. **Sensitivity**: pattern-mixture, selection models, tipping point.

The framing inherits the careful-causal-inference discipline from Chapters 3-4. Missing-data analysis is a special case of causal inference where the

'treatment' is observation status, and the counterfactual is what the value would have been if observed. Many of the same identifying assumptions apply.

10.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) The mechanism is an assumption, not a fact. MAR vs. MNAR cannot be distinguished from the observed data alone. Whatever assumption you make is a substantive claim defended by the design and the context. The biostatistician makes the claim explicit, defends it, and conducts sensitivity analyses to quantify the consequences of violation.

(Judgement 2.) Imputation is modelling. Multiple imputation fits a model for the missing values; that model has assumptions about the joint distribution of the variables. Bad imputation introduces bias more cleanly than complete-case analysis would. The tools make imputation easy; making it correct is the biostatistician's responsibility.

(Judgement 3.) Sensitivity analysis is mandatory. The MAR assumption underlying most modern missing- data methods (multiple imputation, MMRM) cannot be verified from data. Reporting the primary analysis without quantifying what MNAR would do is incomplete. Tipping-point analyses or pattern-mixture models are part of the analysis, not optional.

These judgements distinguish defensible missing-data handling from procedural acceptance of software defaults.

10.4. The three mechanisms

Rubin's classification (Rubin, 1976):

MCAR (Missing Completely At Random). Missingness is independent of all data, observed or unobserved:

$$\Pr(R = 1 \mid Y, X) = \Pr(R = 1).$$

Where R is the missingness indicator. MCAR is testable (compare characteristics of complete and incomplete cases) but rarely true in clinical data.

MAR (Missing At Random). Missingness depends only on observed variables:

$$\Pr(R = 1 \mid Y_{\text{mis}}, Y_{\text{obs}}, X) = \Pr(R = 1 \mid Y_{\text{obs}}, X).$$

Conditional on what we observe, missingness is independent of the unobserved values. The standard assumption underlying multiple imputation, MMRM, and most modern missing-data methods.

MNAR (Missing Not At Random). Missingness depends on the unobserved values:

$$\Pr(R = 1 \mid Y_{\text{mis}}, Y_{\text{obs}}, X) \neq \Pr(R = 1 \mid Y_{\text{obs}}, X).$$

Standard methods are biased; sensitivity analyses are required.

The mechanism is not testable from observed data alone. The defence of MAR is a substantive argument: the variables that drive missingness are observed and included in the imputation model.

10.5. Rubin's framework: multiple imputation

The procedure (Rubin, 1987):

1. Generate M imputed datasets, each with the missing values filled in.
2. Analyse each dataset with the planned model.
3. Combine the M analyses with Rubin's rules: pooled point estimate is the average; pooled variance is the within-imputation variance plus the between-imputation variance plus a finite- sample correction.

Rubin's pooled estimate:

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \theta_m.$$

Rubin's pooled variance:

$$T = \bar{V} + \left(1 + \frac{1}{M}\right) B$$

where $\bar{V}$ is the average within-imputation variance and B is the between-imputation variance.

In R, the `mice` package automates everything:

```
library(mice)

imp <- mice(data, m = 20, seed = 42, printFlag = FALSE)

fit <- with(imp, lm(outcome ~ treatment + age + sex))

pool_fit <- pool(fit)
summary(pool_fit, conf.int = TRUE)
```

The `pool()` function applies Rubin's rules. Output includes pooled point estimate, pooled SE, t-statistic with adjusted degrees of freedom, p-value, 95% CI, fraction of missing information (FMI).

The FMI tells you how much information was lost to missingness; high FMI (above ~30%) means the imputation is contributing substantial uncertainty and the result is sensitive to the imputation model.

10.6. Multiple imputation by chained equations (FCS)

The mechanics:

1. For each variable with missing values, specify a conditional model (predictive mean matching for continuous, logistic for binary, polyreg for nominal).
2. Iterate: impute one variable conditional on the current values of all others; move to the next variable; repeat until convergence.
3. After convergence, generate M imputed datasets.

The `mice` defaults are reasonable starting points:

- Continuous: predictive mean matching (`pmm`). Imputed values are drawn from observed values with similar predicted mean — preserves distributional shape.

- Binary: logistic regression (`logreg`).
- Nominal categorical: multinomial logit (`polyreg`).
- Ordinal: proportional-odds logit (`polr`).

Configure when defaults are wrong:

```
meth <- make.method(data)
meth["age"] <- "pmm"
meth["bmi"] <- "norm"  # normal regression
meth["sex"] <- "logreg"

pred <- make.predictorMatrix(data)
pred[, "id"] <- 0  # do not use id as predictor

imp <- mice(data, method = meth, predictorMatrix = pred,
            m = 20, seed = 42)
```

10.6.1. How many imputations?

The classical rule of thumb $M = 5$ is too few for modern computers. The current recommendation (Buuren, 2018) is M at least equal to the percentage of incomplete cases (e.g., 30 imputations for 30% missing). For trials, $M = 50$ or 100 is common.

10.6.2. The imputation model and the analysis model

Both must be **congenial**: the imputation model should include all the variables in the analysis model, including interactions and any transformations. If the analysis includes $X_1 \times X_2$ interaction, the imputation model should include that interaction term.

Failing congeniality biases the analysis toward 'no interaction' (the imputation does not preserve the relationship that the analysis is trying to estimate). The `mice` package supports including interaction terms in the predictor matrix.

Check your understanding: what to include in the imputation model

Question. A trial has missing data on the primary endpoint and on baseline covariates. The analysis is ANCOVA: outcome ~ treatment + baseline. Should the imputation model include the treatment indicator?

Answer.

Yes. The imputation model must be congenial with the analysis model, which means including all variables the analysis uses. Excluding treatment from the imputation would impose a 'no treatment effect' assumption on the imputation, biasing the analysis toward the null. Including treatment in the imputation lets the imputed values respect the treatment-effect relationship that the analysis is trying to estimate. The same logic applies to all variables in the analysis.

10.7. Sensitivity to MNAR: pattern-mixture models

A pattern-mixture model decomposes the joint distribution by missingness pattern:

$$f(Y) = \sum_{R} f(Y \mid R) \Pr(R).$$

The distribution of Y given missing differs from the distribution of Y given observed; the analyst specifies how.

The standard implementation: **delta-adjustment** imputation. Impute under MAR, then add (or multiply by) a delta to the imputed values to reflect the hypothesised MNAR shift.

```
imp_mar <- mice(data, m = 50, seed = 42)

# delta-adjustment: shift imputed values by 0.5 SD
imp_mnar <- complete(imp_mar, action = "long",
                     include = TRUE)
imp_mnar$outcome[imp_mnar$.imp > 0 & is.na(...)] <-
  imp_mnar$outcome[...] + 0.5 * sd(data$outcome)

# convert back to mids and analyse
```

```
imp_mnar_mids <- as.mids(imp_mnar)
fit_mnar <- with(imp_mnar_mids, lm(outcome ~ treatment))
pool(fit_mnar)
```

The interpretation: 'if we shift the imputed values by 0.5 SD (a worst-case adjustment for the treatment group), does the conclusion change?' Vary delta across a range and report the result for each.

10.8. Selection models

A selection model factorises:

$$f(Y, R) = f(Y) f(R \mid Y).$$

The analyst specifies the missingness model $f(R \mid Y)$ — typically a logit or probit regression of the missingness indicator on the unobserved outcome.

Selection models are computationally heavier than pattern-mixture and require stronger assumptions; in applied trials, pattern-mixture is the more common approach.

10.9. Tipping-point analysis

The tipping-point analysis varies the delta-adjustment across a grid and reports the value at which the primary conclusion changes (the 'tipping point').

```
deltas <- seq(-1, 1, by = 0.1)  # in SD units

results <- map(deltas, function(d) {
  imp_d <- impute_with_delta(imp_mar, d, treatment = 1)
  fit <- with(imp_d, lm(outcome ~ treatment))
  pool(fit)
})
```

```
p_values <- map_dbl(results, ...)

plot(deltas, p_values)
abline(h = 0.05)
```

The plot shows at what delta the p-value crosses 0.05. The tipping point is the answer to 'how strong does the MNAR effect need to be to overturn the conclusion'. The clinical context tells you whether that delta is plausible.

For regulatory trials, the tipping-point analysis is increasingly the expected sensitivity for the primary endpoint. The protocol pre-specifies the grid and the interpretation rule.

10.10. Worked example: missing data in a hypertension trial

The trial from Chapter 8: 800 patients, two arms, 24-week SBP follow-up. About 12% of patients have missing week-24 SBP. The investigator wants to:

1. Run the primary MMRM (handles MAR implicitly).
2. Run a multiple-imputation sensitivity to verify the MMRM result.
3. Run a pattern-mixture sensitivity for MNAR ('jump-to-reference': missing patients in the treatment arm are imputed as if they were in the control arm — a worst-case for the treatment).
4. Run a tipping-point analysis.

```
library(tidyverse)
library(mice)
library(mmrm)

trial <- read_csv("data/bp-trial.csv")

# 1. Primary MMRM
fit_primary <- mmrm(...)
emmeans(fit_primary, ~ treatment | visit,
```

```
        at = list(visit = 24)) |>
  pairs()
# treatment effect: -11.3 mmHg, 95% CI -14.2, -8.4

# 2. Multiple imputation
imp <- mice(trial_long, m = 50, seed = 42,
            method = "pmm")
fit_mi <- with(imp, lm(week24_sbp ~ treatment +
                         baseline_sbp))
pool_mi <- pool(fit_mi)
summary(pool_mi, conf.int = TRUE)
# treatment effect: -11.1 mmHg, 95% CI -14.0, -8.3
# Reassuring: matches MMRM

# 3. Jump-to-reference
trial_jr <- impute_jump_to_reference(
  trial_long, imp,
  group = "treatment",
  reference = 0
)
fit_jr <- with(trial_jr, lm(week24_sbp ~ treatment +
                              baseline_sbp))
pool_jr <- pool(fit_jr)
# treatment effect under JR: -8.7 mmHg, 95% CI -11.5, -5.9
# Still significantly below zero — robust to JR

# 4. Tipping-point
delta_grid <- seq(-2, 0, by = 0.1)  # SD units
tp_results <- map_dfr(delta_grid, function(d) {
  trial_d <- impute_with_delta(trial_long, d,
                               group_for = "treatment")
  fit_d <- with(trial_d, lm(week24_sbp ~ treatment +
                              baseline_sbp))
  pool_d <- pool(fit_d)
  tibble(delta = d, p = summary(pool_d)$p.value[2])
})
```

```
ggplot(tp_results, aes(delta, p)) +
  geom_line() + geom_hline(yintercept = 0.05) +
  labs(x = "Delta (SD units)", y = "p-value")
# tipping point: delta = -1.4 SD
# Interpretation: missing-data shift of 1.4 SD in the
# adverse direction would overturn the conclusion;
# this is implausibly large.
```

The methods section reports each: primary MMRM, multiple imputation as cross-check, jump-to-reference as a stress test, tipping point. The conclusion is that the treatment effect is robust to the missing- data assumption.

10.11. Multiple imputation in time-to-event data

Time-to-event data has its own missing-data structure: covariates may be missing, but the event time is typically observed (or the patient is censored, which is its own kind of 'missing').

Multiple imputation for survival analysis is implemented in `mice` with care to include the Nelson-Aalen estimator of the cumulative hazard as a predictor (so the imputation model knows which patients are at risk). The `smcfcs` package implements substantive-model-compatible FCS for survival outcomes.

The introductory SCAI volume's survival chapter and the SCAI-advanced volume cover the methodology; this chapter notes the issue and points to the specialty references.

10.12. Collaborating with an LLM on missing-data analysis

Three patterns.

Prompt 1: 'Set up multiple imputation for this analysis.' Provide the data structure and the analysis model.

What to watch for. The LLM produces working `mice` code. It often defaults to $M = 5$ (too few) and omits the analysis-model variables from the imputation predictors. Push for M at least equal to the percentage missing and for full congeniality.

Verification. Run with the suggested settings; compare to a more conservative version with $M = 50$. Confirm the analysis-model variables (including interactions) are in the imputation.

Prompt 2: 'Design a sensitivity analysis for unmeasured MNAR in this trial.' Provide the trial setup.

What to watch for. The LLM proposes pattern- mixture or selection. It tends to under-specify the range of delta values and the interpretation rule. Push for a concrete grid and a pre-specified threshold for 'tipping point reached'.

Verification. The grid should span clinically plausible MNAR scenarios. The interpretation should be explicit (delta in original units, in SD units, or in absolute risk change).

Prompt 3: 'Interpret the FMI for this analysis.' Provide the `pool()` output.

What to watch for. The LLM correctly explains FMI. It tends to provide generic thresholds (0.5 = high, etc.) without anchoring to the specific context. Push for the trial-specific implication.

Verification. High FMI (>0.3) is a flag that the result depends substantially on the imputation. Plan for sensitivity.

The meta-pattern: LLMs are good for the syntactic mechanics of `mice` and similar tools and weak at the substantive judgement of how to defend MAR vs. MNAR. Use them for code; bring the substantive reasoning yourself.

10.13. Principle in use

Three habits.

1. **Defend the missing-data mechanism explicitly.** The protocol or SAP states which mechanism is assumed and why.

2. **Run multiple imputation with M at least equal to the percentage of incomplete cases.** Use full congeniality.
3. **Include sensitivity analyses for MNAR.** A tipping-point or pattern-mixture analysis is part of the result; the primary alone is incomplete.

10.14. Exercises

1. Take a dataset with missing values. Compute the missingness pattern (`mice::md.pattern()`). Identify which variables are missing for which subsets.

2. Run `mice` with default settings on a small dataset. Examine the imputed values; check that they look plausible.

3. For a trial with missing primary-endpoint data, run primary MMRM, multiple imputation sensitivity, and jump-to-reference sensitivity. Compare the three estimates.

4. Construct a tipping-point analysis with a grid of delta values. Identify the tipping point and discuss whether it is clinically plausible.

5. For a published trial with reported missing-data handling, identify the assumed mechanism, the primary analysis, and the sensitivity analyses. Comment on whether the sensitivity analyses adequately cover the plausible MNAR scenarios.

10.15. Further reading

- Buuren (2018), *Flexible Imputation of Missing Data* (2nd edition). The reference textbook for the `mice` framework.
- Rubin (1987), *Multiple Imputation for Nonresponse in Surveys*. The foundational text for Rubin's framework.
- National Research Council (2010), the National Research Council report on missing data in clinical trials. Pre-ICH-E9-R1 but still relevant for sensitivity- analysis design.

- The `mice`, `mitml`, and `mi` package documentation are the practical references.

11. Meta-Analysis and Evidence Synthesis

11.1. Learning objectives

By the end of this chapter you should be able to:

- Distinguish fixed-effects, common-effect, and random-effects meta-analysis and choose between them.
- Compute pooled effect estimates and their standard errors using the inverse-variance method, with appropriate handling of heterogeneity.
- Quantify and visualise heterogeneity (τ^2, I^2, H^2) and interpret each.
- Conduct a meta-regression to investigate sources of heterogeneity.
- Produce CONSORT- and PRISMA-compliant reports including forest plots, funnel plots, and Egger's test for publication bias.
- Design a network meta-analysis when multiple treatments are compared across studies.

11.2. Orientation

Meta-analysis is the quantitative synthesis of evidence from multiple studies. The literature is rich; this chapter focuses on what an applied biostatistician needs to conduct one or evaluate one: the core methods (inverse-variance pooling, random effects, heterogeneity quantification), the reporting standards (PRISMA), and the modern extensions (network meta-analysis, individual-patient-data meta-analysis).

The chapter is organised in three threads. **Foundations**: pooling estimators, heterogeneity, forest plots. **Modern methods**: meta-regression, network

meta-analysis, individual-patient-data (IPD) meta-analysis. **Reporting**: PRISMA, GRADE, publication bias.

11.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Pool only what should be pooled. A meta-analysis assumes the included studies estimate a common parameter (or, in random effects, parameters drawn from a common distribution). When studies are clinically too different — different populations, interventions, comparators, outcomes — the pooled estimate is not interpretable as 'the' effect. The biostatistician decides what is poolable; the inclusion criteria and the heterogeneity assessment are the formal record of the decision.

(Judgement 2.) Heterogeneity is a feature, not a nuisance. $I^2 = 75\%$ does not invalidate the meta- analysis; it tells you the studies' results vary beyond chance and that variation is informative. Sources of heterogeneity (population differences, study quality, dose differences) are investigated by meta-regression and subgroup analyses; the pooled estimate is reported alongside the heterogeneity description, not in lieu of it.

(Judgement 3.) Publication bias is real. Studies with statistically significant results are more likely to be published than studies without; this biases meta-analyses toward effect-positive conclusions. The biostatistician investigates with funnel plots, Egger's test, and trim-and-fill, and qualifies the conclusion accordingly. A clean funnel plot does not prove the absence of bias; an asymmetric one is suggestive.

These judgements distinguish a defensible synthesis from a number that combines studies without considering whether they should be combined.

11.4. Inverse-variance pooling

The fundamental method (Borenstein et al., 2009): each study i contributes an estimate $\widehat{\theta}_i$ with variance v_i. The pooled estimate is the inverse-variance-

weighted average:

$$\hat{\theta}_{\text{pool}} = \frac{\sum_i (1/v_i)\hat{\theta}_i}{\sum_i (1/v_i)}$$

with variance:

$$\text{Var}(\hat{\theta}_{\text{pool}}) = \frac{1}{\sum_i (1/v_i)}.$$

This is a **fixed-effects** estimate: it assumes all studies estimate the same parameter, with study-to- study variation being chance.

A **random-effects** estimate adds between-study variance τ^2:

$$\text{Var}_{\text{RE}}(\hat{\theta}_i) = v_i + \tau^2.$$

The pooled estimate uses these inflated variances as weights. With $\tau^2 = 0$, it reduces to fixed effects; with large τ^2, the weights become nearly equal across studies (small studies get more influence).

Estimating τ^2: DerSimonian-Laird (DerSimonian & Laird, 1986) is the classic estimator; restricted maximum likelihood (REML), Paule-Mandel, and Hartung-Knapp adjustments are modern alternatives.

```
library(meta)

ma <- metagen(
  TE = effect_estimate,
  seTE = standard_error,
  studlab = study_name,
  data = study_data,
  sm = "MD",         # mean difference
  fixed = TRUE,
  random = TRUE,
  method.tau = "REML",
  hakn = TRUE        # Hartung-Knapp adjustment
)

summary(ma)
forest(ma)
```

The Hartung-Knapp adjustment (HK or HKSJ) provides better confidence-interval coverage than the standard random-effects approach when the number of studies is small; it is the modern recommendation (IntHout et al., 2014).

11.5. Heterogeneity

Three measures, all derived from Cochran's Q statistic:

τ^2: estimated between-study variance. Has units of the effect (squared); in original units, τ is interpretable as the SD of true effects across studies.

I^2: proportion of total variance attributable to between-study heterogeneity:

$$I^2 = \max\left(0, \frac{Q - df}{Q}\right) \times 100\%.$$

Roughly: 0-25% low, 25-50% moderate, 50-75% substantial, 75-100% considerable. Cited frequently; not the most informative number, since it depends on the sample-size of the included studies.

H^2: ratio of total variance to within-study variance. $H^2 > 1.5$ is typically considered substantial.

The hierarchy of importance: report τ^2 (in original units, with prediction interval), I^2 as a descriptive summary, and H^2 if asked.

A **prediction interval** (Higgins et al., 2009) gives the range in which the true effect of a 'new' study would likely fall:

$$\hat{\theta}_{\text{pool}} \pm t_{n-2}\sqrt{v_{\text{pool}} + \hat{\tau}^2}.$$

The prediction interval is wider than the confidence interval (which is for the pooled mean) and is more informative for clinical interpretation.

11.6. Forest plots

The standard graphical summary:

- One row per study, showing the point estimate and CI as a square (size proportional to the study's weight).
- A diamond at the bottom showing the pooled estimate and its CI.
- Optionally a prediction interval line.

The `meta` and `metafor` packages produce these:

```
forest(ma,
       leftcols = c("studlab", "n.e", "n.c"),
       rightcols = c("effect", "ci"),
       label.left = "Favours treatment",
       label.right = "Favours control",
       prediction = TRUE,
       text.predict = "Prediction interval")
```

The forest plot is the default visual; readers interpret the meta-analysis from it.

11.7. Meta-regression

When heterogeneity is substantial, meta-regression investigates sources. The model:

$$\theta_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, v_i + \tau^2)$$

where X_i is a study-level covariate. The coefficient β_1 tells you how the effect varies with X.

```
mreg <- metareg(ma, ~ baseline_severity)
summary(mreg)
```

Caveat: meta-regression on study-level covariates suffers from **ecological bias**. The relationship between study-level mean baseline severity and the study-level treatment effect is not the same as the relationship between

individual baseline severity and individual treatment effect. Individual-patient- data meta-analysis (below) avoids the issue.

Reserve meta-regression for clearly pre-specified subgroups; do not use it as fishing for explanations of unexplained heterogeneity.

11.8. Network meta-analysis

When multiple treatments are compared (A vs. B, B vs. C, A vs. C across different studies), network meta- analysis estimates the effects of all treatments relative to a reference, using both direct and indirect comparisons.

The `netmeta` package implements frequentist NMA; `gemtc` and `multinma` implement Bayesian NMA.

```
library(netmeta)

nma <- netmeta(TE = effect, seTE = se,
               treat1 = comparator1, treat2 = comparator2,
               studlab = study, data = network_data,
               sm = "MD")
summary(nma)
forest(nma)
netleague(nma)
```

The `netleague` matrix shows pairwise effects for all pairs of treatments. Treatments can also be ranked (SUCRA, P-score) for decision-making.

NMA assumes **transitivity**: the studies comparing different treatments are similar enough that indirect comparisons are valid. The biostatistician evaluates this assumption — it is a substantive judgement based on the included studies' populations and contexts.

11.9. Individual-patient-data meta-analysis

When the original patient-level data from each study is available, IPD meta-analysis provides:

- Standardised analysis (the same model in every study).
- Patient-level subgroup analysis (without ecological bias).
- Better handling of non-linear relationships and interactions.

The cost: data sharing, standardisation, IRB approvals across multiple studies. IPD MA is the gold standard for important questions but is operationally heavy.

Two-stage IPD: fit the model in each study, pool the estimates. One-stage IPD: pool all individual data into one model with random study effects.

11.10. Publication bias

Studies with significant results are more likely to be published. Three diagnostic tools:

Funnel plot: scatter of effect estimate vs. standard error. A symmetric funnel suggests no bias; asymmetry (especially missing studies in the bottom-left, smaller studies with non-significant effects) suggests bias.

```
funnel(ma)
```

Egger's test (Egger et al., 1997): regression-based test for funnel asymmetry. Significant Egger's test suggests bias. Limited power with few studies.

```
metabias(ma, method.bias = "Egger")
```

Trim-and-fill: imputes 'missing' studies from the funnel and recomputes the pooled estimate. The adjustment is a sensitivity, not a definitive correction.

```
ma_tf <- trimfill(ma)
summary(ma_tf)
```

For very few studies (under 10), publication-bias methods have low power; report the funnel plot but qualify the inference.

Check your understanding: random vs. fixed effects

Question. A meta-analysis of 8 RCTs of a hypertension drug shows pooled effect of -3.2 mmHg under fixed effects and -3.5 mmHg under random effects. The two CIs are similar but the random- effects CI is wider. $I^2 = 35\%$. Which estimate should be the primary?
Answer.
Random effects, with Hartung-Knapp adjustment. The fixed-effects model assumes all studies estimate the same parameter, which $I^2 = 35\%$ argues against. Random effects acknowledges the between-study variation and produces a wider, more honest CI. The difference in point estimates (3.2 vs. 3.5) is typically small; the difference in inference (CI width) can be substantial. Modern meta-analyses default to random effects unless there is a strong clinical argument for fixed effects (e.g., very homogeneous studies of the same population). Report the prediction interval alongside the CI for full context.

11.11. PRISMA

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Page et al., 2021) guidelines specify what a systematic review and meta-analysis must report. Highlights:

- **Title**: declares 'systematic review' or 'meta-analysis'.
- **Abstract**: structured (background, methods, results, conclusions).
- **Methods**: the protocol (often pre-registered), inclusion criteria, search strategy, study selection, data extraction, risk of bias, data synthesis, sensitivity analyses.
- **Results**: the PRISMA flow diagram (studies identified, screened, included, excluded), study characteristics table, risk-of-bias assessment, pooled effects, heterogeneity, sensitivity analyses, publication bias.

The `PRISMAstatement` and `prisma2020` R packages generate the flow diagram. Review-quality assessment typically uses Cochrane RoB 2 (RCTs) or ROBINS-I (non-randomised studies).

11.12. GRADE

The Grading of Recommendations, Assessment, Development and Evaluation framework (Guyatt et al., 2008) provides a systematic approach to assessing the quality of evidence in a meta-analysis. GRADE rates the certainty of evidence (high, moderate, low, very low) based on:

- Risk of bias in the included studies.
- Inconsistency (heterogeneity).
- Indirectness (population, intervention, comparator, outcome match).
- Imprecision (CI width).
- Publication bias.

The GRADEpro tool implements the framework. Modern systematic reviews are increasingly expected to include GRADE assessments.

11.13. Worked example: a meta-analysis of SGLT2 trials

Eight RCTs of SGLT2 inhibitors in heart failure with reduced ejection fraction (HFrEF). Outcome: all-cause mortality at 12 months. Each trial reports an HR with 95% CI.

```
library(tidyverse)
library(meta)

# Data: one row per trial
sglt2_meta <- tribble(
  ~study,           ~hr,    ~lower,  ~upper, ~n_t, ~n_c,
  "DAPA-HF",        0.83,   0.71,    0.97,   2373, 2371,
  "EMPEROR-Reduced", 0.92,   0.77,    1.10,   1863, 1867,
  "DEFINE-HF",      0.79,   0.65,    0.96,   1234, 1230,
  # ... etc
)

# convert HR to log-HR for pooling
```

```
sglt2_meta <- sglt2_meta |>
  mutate(loghr = log(hr),
         se_loghr = (log(upper) - log(lower)) / (2 * 1.96))

# random-effects meta-analysis
ma <- metagen(TE = loghr, seTE = se_loghr,
              studlab = study,
              data = sglt2_meta,
              sm = "HR",
              method.tau = "REML",
              hakn = TRUE)
summary(ma)
# pooled HR: 0.86 (95% CI 0.79-0.94, p < 0.001)
# I^2 = 18%, tau^2 = 0.005

# forest plot
forest(ma)

# funnel and Egger
funnel(ma)
metabias(ma, method.bias = "Egger")
# p = 0.42, no evidence of asymmetry

# meta-regression on baseline NT-proBNP
mreg <- metareg(ma, ~ baseline_ntprobnp)
summary(mreg)
# beta_1 = ..., not significant
```

The reported synthesis: 'Across 8 RCTs (N = ~22,000), SGLT2 inhibitors reduced 12-month all-cause mortality by 14% (pooled HR 0.86, 95% CI 0.79-0.94) compared with placebo. Heterogeneity was low (I^2 = 18%; τ = 0.07). Meta-regression on baseline NT-proBNP did not explain the residual heterogeneity. Egger's test showed no evidence of publication bias. The result was consistent across SGLT2 agents.'

11.14. Collaborating with an LLM on meta-analysis

Three patterns.

Prompt 1: 'Pool these studies into a meta-analysis.' Provide the trial-level data.

What to watch for. The LLM produces working `meta` or `metafor` code. It often defaults to fixed-effects or DL random-effects without HK adjustment. Push for HK and prediction interval.

Verification. Run with the suggested settings; verify τ^2, I^2, and the prediction-interval width match the data.

Prompt 2: 'What does this $I^2 = 65\%$ tell me?' Provide the meta-analysis output.

What to watch for. The LLM correctly explains I^2 but tends to focus on the descriptive label (high heterogeneity). Push for the implications: prediction interval will be wide, meta-regression may be warranted, the pooled mean is one summary among several.

Verification. Read the LLM's interpretation against the prediction interval. The width of the PI is the practical implication.

Prompt 3: 'Diagnose publication bias in this meta-analysis.' Provide the funnel plot, Egger's test result.

What to watch for. The LLM produces a competent discussion. With few studies (under 10), it may overstate the power of Egger's test. Push for the appropriate qualification.

Verification. The funnel plot is the visual; the Egger's test is supportive. Both should be reported, and the interpretation should be qualified by the number of studies.

The meta-pattern: LLMs are good for the syntactic mechanics of meta-analysis (writing the `metagen` call, drafting the methods) and weak at the substantive judgement (whether the studies are poolable, whether heterogeneity is concerning). Use them for code, bring substantive judgement yourself.

11.15. Principle in use

Three habits.

1. **Random effects with HK as the default.** Fixed effects only when the studies are unequivocally homogeneous.

2. **Report the prediction interval.** It is the single most informative summary of how variable true effects are likely to be in a new setting.

3. **Address publication bias explicitly.** Funnel plot, Egger's (with appropriate caveats), trim-and-fill as sensitivity. With few studies, the report names the limit.

11.16. Exercises

1. Take a published meta-analysis in your field. Reproduce the pooled effect from the trial-level data; verify I^2, τ^2, and CI.

2. Compute the prediction interval for a published meta-analysis. Compare the width to the CI for the pooled mean. Discuss the practical implication.

3. For a meta-analysis with $I^2 = 70\%$, conduct a meta-regression on a candidate moderator. Discuss whether the moderator explains the heterogeneity.

4. Construct a funnel plot for a meta-analysis of 8 studies. Apply Egger's test and trim-and-fill. Report the result.

5. Design a hypothetical network meta-analysis with four treatments compared across 6 studies. Identify the direct and indirect comparisons and the transitivity assumption that must hold.

11.17. Further reading

- Borenstein et al. (2009), *Introduction to Meta-Analysis*. The applied textbook.
- Higgins et al. (2019), *Cochrane Handbook for Systematic Reviews of Interventions*. The reference for systematic-review methodology.
- Rothstein et al. (2005), *Publication Bias in Meta-Analysis*. The reference treatment.
- Page et al. (2021), the PRISMA 2020 statement and flow diagram. The reporting standard.
- The `meta`, `metafor`, `netmeta`, and `multinma` R packages are the practical tools.

12. Categorical Data, Advanced

12.1. Learning objectives

By the end of this chapter you should be able to:

- Fit ordinal regression with the proportional-odds model and recognise when the proportional-odds assumption is violated.
- Fit multinomial regression for nominal categorical outcomes and interpret the relative-risk-ratio output.
- Fit log-linear models for multi-way contingency tables and use them for tests of independence and conditional independence.
- Apply exact methods (Fisher, conditional logistic for matched data) when sample sizes are small or cells sparse.
- Compute and interpret agreement and reliability measures (Cohen's kappa, weighted kappa, ICC) for categorical data.
- Evaluate diagnostic-test performance with sensitivity, specificity, predictive values, and ROC curves.

12.2. Orientation

The introductory volume's GLM chapter covered logistic regression for binary outcomes. This chapter extends to the broader categorical-data toolkit: ordinal, multinomial, log-linear, exact methods, agreement, and diagnostic-test evaluation. The material is useful across clinical research, epidemiology, and outcomes research; this chapter makes it production-ready.

The chapter is organised in four threads. **Beyond binary**: ordinal and multinomial regression. **Multi-way tables**: log-linear models, exact meth-

ods. **Agreement**: kappa, ICC. **Diagnostic tests**: sensitivity, specificity, predictive values, ROC.

12.3. The statistician's contribution

Three judgements are not delegable.

(Judgement 1.) Match the model to the outcome's structure. A binary outcome is logistic. An ordinal outcome (low / medium / high) loses information when treated as binary or as numeric; ordinal regression is the right tool. A nominal outcome (red / green / blue) cannot be ordered; multinomial regression is the right tool. The biostatistician identifies the outcome's structure and chooses the model accordingly.

(Judgement 2.) Test the proportional-odds assumption. Ordinal regression assumes the covariate's effect is the same across all splits of the ordinal variable. The assumption is testable; the partial-proportional-odds and generalised-ordinal alternatives address violations. Skipping the test can produce misleading inference.

(Judgement 3.) Sparse tables demand exact methods. A 2x2 table with any expected count under 5 is the textbook case for Fisher's exact. The chi-squared p-value is unreliable. Modern computing makes exact methods cheap; use them when the asymptotic approximation is suspect.

These judgements distinguish a categorical-data analysis that respects the data's structure from one that loses information or trusts unreliable p-values.

12.4. Ordinal regression: the proportional-odds model

For an ordinal outcome with categories $1, 2, \dots, K$, the proportional-odds model (McCullagh, 1980):

$$\log \frac{\Pr(Y \le k)}{\Pr(Y > k)} = \alpha_k + X^T \beta, \quad k = 1, \dots, K-1.$$

The α_k are category-specific intercepts; the β is the same across categories — that is the proportional-odds (PO) assumption. The interpretation of β_j: a one-unit change in X_j multiplies the cumulative odds (of being in category $\leq k$ vs. $> k$) by $\exp(\beta_j)$, regardless of which k.

Implementation:

```
library(MASS)

fit <- polr(disease_severity ~ age + sex + treatment,
            data = trial, Hess = TRUE)
summary(fit)
exp(coef(fit))   # odds ratios
```

`Hess = TRUE` requests the Hessian, needed for standard errors.

The `ordinal` package provides more flexibility (mixed effects, alternative link functions).

12.4.1. Testing proportional odds

```
library(brant)
brant(fit)   # Brant test for PO assumption
```

A significant Brant test indicates PO violation. Two responses:

- **Partial proportional odds**: relax PO for the violating covariates. Implemented in `VGAM::vglm()`.
- **Generalised ordinal regression** (alternative parameterisations): allow effects to differ across splits.

For non-proportional ordinal data, multinomial regression is also an option (loses the ordering but fits the data).

12.5. Multinomial regression

For a nominal outcome with K categories, multinomial logistic regression models each category against a reference:

$$\log \frac{\Pr(Y = k)}{\Pr(Y = K)} = \alpha_k + X^T \beta_k, \quad k = 1, \dots, K-1.$$

Each non-reference category has its own coefficient vector β_k. The interpretation of $\exp(\beta_{kj})$: the relative risk of category k vs. the reference, per unit change in X_j, holding others fixed.

```
library(nnet)

fit <- multinom(treatment_choice ~ age + sex +
                  comorbidity, data = patients)
summary(fit)
exp(coef(fit))    # relative risk ratios
```

Multinomial output is harder to interpret than binary or ordinal because there are $K-1$ vectors of coefficients. The clearest reporting often uses predicted probabilities at specific covariate values rather than the coefficients themselves; the `marginaleffects` package handles this cleanly.

12.6. Log-linear models for multi-way tables

When you have a multi-way contingency table (say, sex × treatment × outcome), the log-linear model parameterises the cell counts:

$$\log \mu_{ijk} = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk}.$$

Different combinations of the interaction terms correspond to different conditional-independence structures. For example:

- All single terms only: complete independence.
- All two-way interactions, no three-way: conditional independence given pairs.

```
library(MASS)

tab <- table(patients$sex, patients$treatment,
             patients$outcome)
fit <- loglm(~ sex * treatment + treatment * outcome +
                sex * outcome,
             data = tab)
summary(fit)
```

The model fits a structure (in this case, two-way interactions but no three-way), and the test is whether the data are consistent with that structure.

Log-linear models generalise the chi-squared test of independence to more than two variables. For most applied work, the simpler logistic-regression formulation (with one variable as outcome) is more interpretable; log-linear models are the right tool when no variable is naturally the outcome.

12.7. Exact methods

The chi-squared test approximates the sampling distribution of the test statistic under the null. The approximation is good when expected counts are large (textbook rule: all expected 5). For smaller tables or sparse cells, exact methods compute the distribution directly.

Fisher's exact test for 2x2 tables:

```
fisher.test(table(d$exposed, d$outcome))
```

The test conditions on the row and column margins and computes the exact p-value. For larger tables, the exact computation is slow but feasible with `fisher.test(..., simulate.p.value = TRUE)`.

Conditional logistic regression for matched case-control studies (where matching is on a variable that would be a confounder):

```
library(survival)

fit_clogit <- clogit(case ~ exposure + age + strata(matched_set),
                     data = matched_data)
summary(fit_clogit)
```

The conditional likelihood eliminates the matched-set intercepts, providing valid inference when there are many matched sets.

12.8. Agreement and reliability

When two raters classify the same items, **Cohen's kappa** measures agreement beyond chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed proportion of agreement and p_e is the expected agreement under independence. Range: -1 to 1. Conventional benchmarks:

- 0.0–0.20: slight
- 0.21–0.40: fair
- 0.41–0.60: moderate
- 0.61–0.80: substantial
- 0.81–1.00: almost perfect

```
library(irr)

kappa2(cbind(rater1, rater2))
```

For ordinal scales, **weighted kappa** (linear or quadratic weights) gives more credit for near- agreements than far ones. Quadratic weighted kappa is the standard for ordinal data; on a 5-point scale it is approximately equivalent to the ICC for the same data.

For continuous data, the **intraclass correlation coefficient (ICC)** is the analogous measure:

```
library(psych)

ICC(cbind(rater1, rater2))
```

The ICC has variants: ICC(1,1), ICC(2,1), ICC(3,1), each with single-rater or average-rater forms. The choice depends on whether raters are fixed or random and whether the analysis is reporting agreement of single ratings or means of multiple ratings.

12.9. Diagnostic-test evaluation

For a binary diagnostic test:

- **Sensitivity** = Pr(positive | disease). How often the test catches actual cases.
- **Specificity** = Pr(negative | healthy). How often the test correctly rules out non-cases.
- **Positive predictive value (PPV)** = Pr(disease | positive). Of those testing positive, what fraction are actually diseased.
- **Negative predictive value (NPV)** = Pr(healthy | negative).

PPV and NPV depend on disease prevalence; sensitivity and specificity do not. A test with 99% sensitivity and 99% specificity has poor PPV in a low-prevalence population (one false positive per true positive at 1% prevalence).

For a continuous test (a biomarker, a risk score), the **ROC curve** plots sensitivity vs. 1-specificity as the threshold varies. The **AUC** is the area under the curve; AUC = 0.5 is no better than chance, AUC = 1 is perfect.

```
library(pROC)

roc_obj <- roc(disease_status, biomarker_value)
plot(roc_obj)
auc(roc_obj)
```

Choose a threshold based on the application: maximise the Youden index (sensitivity + specificity - 1), constrain by required sensitivity, or weight by the costs of false positives and false negatives.

For multiple tests, the **DeLong test** compares ROC curves:

```
roc.test(roc1, roc2)
```

Check your understanding: when prevalence matters

Question. A new screening test for a rare disease has 99% sensitivity and 95% specificity. Disease prevalence is 1 in 1000. What is the PPV?
Answer.
Apply Bayes:

$$\text{PPV} = \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot 0.999} = 0.0194.$$

The PPV is just 1.94%. About 50 people test positive for every actual case. The test, despite excellent sensitivity and specificity, is not useful as a standalone screening tool in this population because the false-positive rate dominates the true-positive rate. This is why screening tests are often hierarchical (a sensitive first test followed by a specific confirmation) rather than single tests.

12.10. Calibration of risk scores

Beyond discrimination (AUC), a risk score should be **calibrated**: the predicted probability should match the observed risk. A model that says '20% risk' for a group should produce 20% events in that group.

Calibration plot: deciles of predicted probability on the x-axis, observed event rate on the y-axis. The line should fall on the diagonal.

```
library(rms)
```

```
cal <- val.prob(predicted_prob, actual_outcome,
                pl = TRUE)
```

A model with high AUC but poor calibration is over- or under-confident; **recalibration** (Platt scaling, isotonic) corrects it.

For a clinical decision-support tool that uses absolute risk thresholds (treat above 10%, do not treat below 5%), calibration is the binding constraint; discrimination alone is insufficient.

12.11. Worked example: a multi-faceted categorical analysis

A study evaluates a new diagnostic test for a condition. The test has continuous output (a biomarker score). The reference standard is a definitive (but expensive) test. The analyst wants to:

1. Compute sensitivity, specificity, PPV, NPV at a chosen threshold.
2. Plot the ROC curve and compute AUC.
3. Compare to an existing test (DeLong test).
4. Examine calibration of the new test as a risk score.
5. Conduct an inter-rater reliability analysis (two technicians independently scoring the test).

```
library(pROC)
library(rms)
library(irr)

study <- read_csv("data/diagnostic-test.csv")

# 1. Sensitivity, specificity, etc.
predicted <- study$new_test_score > 50
truth <- study$true_disease

confusion <- table(predicted, truth)
sens <- confusion[2,2] / sum(confusion[,2])
```

```
spec <- confusion[1,1] / sum(confusion[,1])
ppv <- confusion[2,2] / sum(confusion[2,])
npv <- confusion[1,1] / sum(confusion[1,])
sens; spec; ppv; npv

# 2. ROC curve
roc_new <- roc(truth, study$new_test_score)
plot(roc_new)
auc(roc_new)
# 0.91

# 3. DeLong vs. existing test
roc_old <- roc(truth, study$old_test_score)
roc.test(roc_new, roc_old)
# z = 4.2, p < 0.001 — the new test is significantly better

# 4. Calibration as risk score
val.prob(study$new_test_predicted_prob, truth, pl = TRUE)
# slope 0.95, intercept 0.02 — well calibrated

# 5. Inter-rater reliability
inter_rater <- study |>
  filter(both_raters == 1) |>
  select(rater1_score, rater2_score)
icc(inter_rater)
# ICC = 0.89 — substantial agreement
```

The reported analysis: 'The new test had AUC 0.91 (95% CI 0.87-0.95), significantly better than the comparator test (AUC 0.79, p < 0.001 by DeLong). At the proposed threshold of 50, sensitivity was 87% (95% CI 80-93%), specificity 88%, PPV 64%, NPV 96%. The new test was well-calibrated as a risk score (slope 0.95, intercept 0.02). Inter-rater reliability was substantial (ICC 0.89, 95% CI 0.85-0.92).'

12.12. Collaborating with an LLM on advanced categorical-data analysis

Three patterns.

Prompt 1: 'Fit the right model for this categorical outcome.' Provide the data and the question.

What to watch for. The LLM correctly distinguishes ordinal from nominal but sometimes defaults to multinomial when ordinal would preserve information. Push for the proportional-odds test on ordinal outcomes.

Verification. Use the Brant test or fit both ordinal and multinomial; compare.

Prompt 2: 'Choose the threshold for this diagnostic test.' Provide the ROC curve data and the clinical context.

What to watch for. The LLM defaults to maximising Youden index. The right threshold depends on the costs of false positives vs. false negatives in the specific clinical context. Push for the substantive trade-off.

Verification. The threshold is a clinical decision informed by the costs. The LLM provides the analytical machinery; the substantive judgement is yours.

Prompt 3: 'Compute kappa for this rater agreement.' Provide the ratings.

What to watch for. The LLM correctly computes unweighted kappa. For ordinal scales, push for quadratic weighting.

Verification. The choice of weighting matters more than the number; check the literature for what the field's convention is.

The meta-pattern: LLMs are good at the syntactic mechanics (writing the model call) and weak at the substantive judgement (which model fits the question, which threshold matches the clinical trade-off). Use them for code; bring substantive judgement yourself.

12.13. Principle in use

Three habits.

1. **Match the model to the outcome's structure.** Ordinal models for ordinal data; multinomial for nominal; log-linear for multi-way without an outcome.

2. **Test the proportional-odds assumption.** Brant or partial-PO if needed; if PO fails substantially, consider multinomial.

3. **Pair discrimination with calibration.** AUC alone is insufficient for any model used as a risk score. Calibration must be reported.

12.14. Exercises

1. For an ordinal outcome (e.g., a 5-point disease severity scale), fit a proportional-odds model. Test the PO assumption; if it fails, fit a partial-PO alternative and compare.

2. For a multinomial outcome (e.g., one of three treatment choices), fit multinomial logistic regression. Report relative risk ratios for two covariates and interpret them.

3. For a 2x2 table with one cell count of 2, run both chi-squared and Fisher's exact. Compare the p-values.

4. For a diagnostic test, compute sensitivity, specificity, PPV, and NPV at three thresholds. Plot the ROC curve and compute AUC.

5. For an inter-rater reliability study with two raters and an ordinal scale, compute Cohen's kappa, weighted kappa with linear weights, and weighted kappa with quadratic weights. Interpret each.

12.15. Further reading

- Agresti (2013), *Categorical Data Analysis* (3rd edition). The reference textbook.
- McCullagh (1980), 'Regression models for ordinal data'. The foundational ordinal-regression paper.
- Harrell (2015), *Regression Modeling Strategies*. The reference for risk-score calibration.
- Pepe (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*. The reference for diagnostic-test methodology.
- The `MASS`, `nnet`, `ordinal`, `pROC`, `rms`, and `irr` packages are the practical tools.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, *133*(6), 601–609.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–972.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Buuren, S. van. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman; Hall/CRC. https://stefvanbuuren.name/fimd/

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Committee for Proprietary Medicinal Products. (2002). *Points to consider on multiplicity issues in clinical trials* (CPMP/EWP/908/99). https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials-scientific-guideline

Conley, T. G., Hansen, C. B., & Rossi, P. E. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, *94*(1), 260–272.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford University Press.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Wiley.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2015). *Fundamentals of clinical trials* (5th ed.). Springer.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer.

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, *108*(5), 616–619.

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman; Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A*, *172*(1), 137–159.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

International Council for Harmonisation. (2019). *ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

IntHout, J., Ioannidis, J. P. A., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird

method. *BMC Medical Research Methodology*, *14*, 25.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman; Hall/CRC.

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). Springer.

Knol, M. J., Groenwold, R. H. H., & Grobbee, D. E. (2012). P-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. *European Journal of Preventive Cardiology*, *19*(2), 231–232.

Laan, M. J. van der, & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1).

Lash, T. L., VanderWeele, T. J., Haneuse, S., & Rothman, K. J. (2021). *Modern epidemiology* (4th ed.). Wolters Kluwer.

Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., & Zwarenstein, M. (2015). The PRECIS-2 tool: Designing trials that are fit for purpose. *BMJ*, *350*, h2147.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, *42*(2), 109–142.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. The National Academies Press. https://doi.org/10.17226/12955

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science (Translated Reprint)*, *5*, 465–472.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Piantadosi, S. (2017). *Clinical trials: A methodologic perspective* (3rd ed.). Wiley.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman; Hall/CRC.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9-12), 1393–1512.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Schulz, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c332.

Shi, B., Choirat, C., Coull, B. A., VanderWeele, T. J., & Valeri, L. (2021). CMAverse: A suite of functions for reproducible causal mediation analyses. *Epidemiology*, *32*(5), e20–e22.

Taves, D. R. (2010). The use of minimization in clinical trials. *Contemporary Clinical Trials*, *31*(2), 180–184.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., & Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, *32*(22), 2380–2385.

U.S. Food and Drug Administration. (2010). *Guidance for the use of Bayesian statistics in medical device clinical trials*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials

U.S. Food and Drug Administration. (2019). *Adaptive designs for clinical trials of drugs and biologics: Guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-

design-clinical-trials-drugs-and-biologics-guidance-industry

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*(4), 268–274.

Credits

The chapter list was constructed by surveying 12 publicly-available syllabi from major US biostatistics programmes (documented in `docs/syllabi-survey.md`) and adopting the topics that appeared in three or more of them.

Cover artwork generated procedurally by the Python script at `images/build-cover.py`. Watercolour palette: deep slate through silver-blue to warm gold and pale cream, anchored on a brand slate blue-grey `#4a5a6c`. Typography: Avenir family (Apple system font); body text in Source Serif 4.

Colophon

This book was produced with Quarto, typeset in Source Serif 4, with code blocks set in JetBrains Mono. Source code is in R 4.4 or later.

The book is hosted at https://scai-advanced.rgtlab.org on Netlify, with continuous deployment via GitHub Actions from the `rgt47/scai-advanced` repository on every push to `main`. The deployment recipe is in `HOSTING.md`.

The cover is generated procedurally by `images/build-cover.py` from a watercolour gradient anchored on `#7a2c4e` (the volume's brand burgundy) and overlaid with Avenir typography. Re-running the script regenerates the cover; the procedural approach makes the cover reproducible and editable in version control rather than dependent on a one-shot AI image-generation step.

Last rendered: see the build timestamp on the homepage.