Table of Contents

OrthoScienus Teaching · Research Skills Course 1

Module 01 · Formulating the answerable question 2

Module 02 · Study design selection 4

Module 03 · Bias control 6

Module 04 · Outcomes and measurement 8

Module 05 · Sample size and power 10

Module 06 · Data management 12

Module 07 · Analysis and estimation 14

Module 08 · Causal inference from observational data 16

Module 09 · Systematic review workflow 18

Module 10 · Risk of bias and certainty of evidence 20

Module 11 · Reporting guidelines 21

Module 12 · Reproducibility and workflow 23

Module 13 · Scientific writing 25

Module 14 · Data visualization 27

Module 15 · Peer review 30

Module 16 · Presentation and risk communication 33

Module 17 · Research ethics and integrity 35

Module 18 · Calibration, decision journals, and deliberate practice 37

Full reference list — JOA style 39

Foundational works not cited above but worth owning 45

A closing note on this course 46

OrthoScienus Teaching · Research Skills Course

Clear thinking. Honest evidence. Better decisions.

A practical course in clinical research — from question to publication to podium — built for surgeons who want reproducible workflows, not slogans. Arthroplasty-flavored examples. Checklist-first. Reference-anchored.

Author · Jack Walsh, DO · Arthroplasty Surgeon-Scientist · Primary and revision hip and knee replacement

How to use this course

Each module has the same spine: **Concept · Checklist · Worked example · Pitfalls · Self-check · Anchor references.** Read the concept once, keep the checklist beside your project, and return to the worked example when your own study looks similar. Use the self-check to force retrieval before you look at the answer — that is where the learning is. References follow numbered JOA style; the full bibliography is at the end.

How this connects to the OrthoScienus framework

- Modules 01–08 feed the **Evidence Engine** — the controlled path from question to reviewable output.
- Modules 05, 07, 08, 14 feed **Estimation First** — magnitude and precision before threshold labels.
- Modules 09, 10, 11, 17 feed the **Verification Layer** — reporting standards, guideline appraisal, and integrity checks that gate submission and adoption.
- Modules 12, 13, 15, 16, 18 feed the **Decision Journal** — reproducibility, writing, review, communication, prediction, outcome, debrief, update.

A caveat you should not skip. This is a curated survey, not a systematic review. Anchor references were selected because they teach a step well, not because they are the only or best paper on their topic. The Cochrane Handbook, the EQUATOR Network, Rothman's *Modern Epidemiology*, Hernán and Robins' *Causal Inference: What If*, and the ICMJE Recommendations sit beneath this course and are worth owning outright.

Module 01 · Formulating the answerable question

Learning objectives. After this module you can (a) convert a clinical hunch into a structured, answerable research question; (b) apply PICO (or PECO/PICOTS) to search PubMed effectively; (c) apply FINER to test whether the question is worth doing at all.

Concept — brief. A good research question is answerable, feasible, and worth answering. Most projects fail here, quietly, months later. PICO — Population, Intervention, Comparator, Outcome — is the working structure for intervention questions and the search-engine grammar of PubMed [1]. FINER — Feasible, Interesting, Novel, Ethical, Relevant — is the sanity check on whether to spend a year of your life on it [2,3].

The question dictates the design. If you cannot state your comparator in one sentence, you do not have an answerable question yet. If you cannot state your outcome as a measurable event with a time window, you have a topic, not a study.

Checklist — question formulation

- ☐ Population is defined by inclusion **and** exclusion criteria, with the case-mix that matches the decision you want to inform.

- ☐ Intervention is described in enough detail that another surgeon could replicate the exposure (implant, approach, technique, adjunct).
- ☐ Comparator is a real, plausible alternative — not "no treatment" when no treatment isn't offered.
- ☐ Outcome is a specific instrument, event, or measurement, with a time point.
- ☐ Timing (T) is specified — index event → outcome window.
- ☐ Setting (S) is specified — single center vs. registry vs. multi-center; academic vs. community.
- ☐ FINER passed: Feasible with your data, Interesting to a decision-maker, Novel relative to the last 3 years of literature, Ethical, Relevant to a clinical decision or policy.

Worked example — arthroplasty.

Rough hunch: "Are cementless stems doing better in older women now?"

Structured question: *In women aged 70 or older undergoing elective primary THA for osteoarthritis (**P**), does a modern taper-wedge cementless femoral stem (**I**) compared to a polished tapered cemented stem (**C**) result in a lower cumulative incidence of periprosthetic femoral fracture (**O**) at 2 years post-op (**T**), using national registry data (**S**)?*

FINER check: Feasible if you have registry access; Interesting because current guidance still favors cement in this group; Novel because most data pre-date modern taper-wedge geometries; Ethical (observational); Relevant to intraoperative decision-making. Passes.

Pitfalls

- **"How does X compare?"** with no comparator specified is a topic, not a question.
- **Composite outcomes** ("any complication") hide effect sizes and inflate event counts. Prefer a single primary outcome with prespecified secondaries.
- **Registry-defined populations drift** — the case-mix of women 70+ in a registry is not the case-mix in your clinic.
- **"Novel" via PubMed absence** is not novel. Search the last 3 years of top ortho journals plus general medical journals; if the question is important, someone probably asked it.

Self-check

1. Why does 'PICO' fail as a search grammar when the comparator is 'usual care'?
2. You have a PubMed search returning 12 hits. Should you feel confident the question is novel?
3. A colleague proposes: 'Does robotic TKA improve outcomes?' What are the two structural defects in this question?

Answers

1. Because ‘usual care’ varies by setting and rarely maps to a clean PubMed MeSH term; you end up retrieving heterogeneous comparators and diluting the signal. Fix: specify the actual usual-care components (e.g., ‘cemented tapered stem’) or use PICOTS and constrain by setting.

2. No. Twelve hits from a single PubMed pass is not novelty evidence — it is a search-strategy artefact. Check Embase, Cochrane CENTRAL, and the last 3 years of the top 3 journals in the field before claiming novelty.

3. First: No comparator specified (robotic vs. what — conventional jig, navigation, or PSI?). Second: ‘Outcomes’ is a category, not an outcome — no instrument, no time point, no MCID reference. Rewrite as a PICOTS question with a single primary outcome and time window.

Anchor references [1,2,3]

Module 02 · Study design selection

Learning objectives. After this module you can match a question to the strongest design that is feasible; explain when a pragmatic trial or target-trial-emulated cohort is the honest answer; recognize when your design cannot answer the question you posed.

Concept — brief. The hierarchy that gets taught (RCT > cohort > case-control) is a useful default and a dangerous crutch. A well-designed cohort can beat a poorly executed RCT. The right question is: *what design would answer this question under ideal conditions, and how closely can I approximate it?* That is the essence of the **target trial** framework — you specify the RCT you would run if you could, then either run it, emulate it observationally, or state clearly why you cannot [4,5].

Pragmatic trials sit between explanatory RCTs and real-world observational studies. They test whether an intervention works in usual care, with usual patients, at usual sites [6]. Their price is external validity purchased with less internal control.

Checklist — design selection

- ☐ Write the target trial protocol first — eligibility, treatment strategies, assignment, follow-up, outcome, causal contrast.
- ☐ Ask which components you can actually implement (randomization, blinding, follow-up completeness).
- ☐ Where you must deviate, name the deviation and its expected bias.
- ☐ Prespecify a **primary analysis** that respects the design (intention-to-treat for RCT; time-zero-aligned per-protocol for target trial emulation).
- ☐ Register the protocol (Module 12) before enrolling or extracting data.

Design-to-question map — rough

Question type	First-choice design	Watch-out
Efficacy of a single implant modification	Small explanatory RCT	Underpowered for rare complications
Effectiveness at scale	Pragmatic RCT or registry cohort with TTE	Cluster effects, unmeasured confounding
Rare adverse event	Registry cohort or nested case-control	Ascertainment bias
Diagnostic test performance	Cross-sectional with reference standard	Spectrum bias
Prognosis	Prospective cohort	Attrition, competing risks
Mechanism	Bench, cadaver, or physiologic study	Limited external validity

Worked example — arthroplasty.

Question: *Does routine tranexamic acid reduce transfusion in primary TKA?* A large well-blinded RCT already exists, so the target trial is trivially "the trial that was done." Your incremental contribution is not another small RCT — it is either a subgroup you can address in a registry TTE (e.g., patients on chronic anticoagulation, often excluded from RCTs) or an implementation study (does a hospital-level protocol change transfusion rates in usual care?). The design changes because the question changes.

Pitfalls

- **Immortal time bias** — defining exposure using post-baseline information. The target trial framework prevents it by aligning time zero, eligibility, and treatment assignment [5].
- **Ecologic confusion** — inferring individual-level effects from aggregate data.
- **"We'll just adjust for it"** is not a substitute for design. Adjustment cannot fix unmeasured confounding or selection.

Self-check

1. You have registry data on 40,000 THAs and want to compare two cementless stems. Why is 'target trial emulation' the right frame, and what is the first artifact you produce?
2. A reviewer says 'this is only a retrospective cohort; it cannot answer the question.' When is that reviewer wrong?
3. Give one example where a pragmatic RCT is the wrong choice.

Answers

1. TTE keeps you honest about time zero, eligibility, and treatment assignment — the three places observational analyses drift into immortal-time bias. First artifact: the

target-trial protocol document, written as if you were running the RCT, including the causal contrast.

2. When the target trial you would run has been done and is unlikely to be repeated (ethics, cost, equipoise), and your cohort is a well-executed emulation with prespecified time zero and a defensible confounder set. Design hierarchy is a default heuristic, not a rule.

3. A first-in-human dose-finding study — internal validity dominates and generalizability is premature. Pragmatic trials trade internal control for external validity; use them once efficacy is established.

Anchor references [4,5,6]

Module 03 · Bias control

Learning objectives. After this module you can name the major biases that threaten your study *before* it runs; deploy specific design and analysis moves to blunt each; distinguish confounding, selection bias, and information bias.

Concept — brief. Bias is not a moral failing. It is a systematic error introduced by the design, the setting, the measurement, or the analysis. Three families dominate:

- **Selection bias** — who ends up in your sample is related to both exposure and outcome.
- **Information bias** — how exposure or outcome is measured differs by group.
- **Confounding** — a third variable causes both exposure and outcome.

Randomization addresses confounding by design. Blinding addresses information bias and interpretation bias. Complete follow-up addresses attrition/selection during the study. Careful sampling addresses selection at study entry. Confounder adjustment (Module 08) addresses residual confounding in observational studies.

Even in the analysis phase, **blinded interpretation of results** — deciding what your findings mean before you look at the effect direction — reduces confirmation bias and post-hoc rationalization [7].

Checklist — bias control

- ☐ Selection: eligibility criteria are applied identically to potential cases in exposed and unexposed groups.
- ☐ Information: outcome ascertainment uses the same instrument, threshold, and rater across groups; assessor blinded where feasible.
- ☐ Confounding: baseline covariates specified *a priori* using a DAG (Module 08), not "everything we measured."

- ☐ Collider bias: draw the causes of each proposed adjustment variable. If exposure and another cause of the outcome both affect it, adjusting for it can open a noncausal path [8].
- ☐ For observational studies, a **sensitivity analysis to unmeasured confounding** is planned (e.g., E-value) [9].
- ☐ Blinded interpretation: the analytic plan states what conclusion each result *pattern* would support, written before unblinding.

Worked example — arthroplasty.

You want to compare 90-day readmission after THA between two approaches (posterior vs. anterior). The surgeons who do anterior are also, on average, higher-volume. If you adjust only for age/BMI/ASA, you have adjusted for the wrong thing — surgeon volume is a confounder. Suppose the approach affects length of stay, which in turn affects readmission. Length of stay is then a **mediator**. Adjusting for it blocks that part of the total effect. Draw the DAG before you run the regression.

Pitfalls

- **Table 1 p-values** are not a bias check. They tell you nothing about which variables are confounders.
- **Propensity scores** balance measured confounders only. They do nothing for unmeasured ones.
- **Complete-case analysis** silently selects the sample if missingness relates to outcome.
- **Post-hoc subgroup analyses** as "sensitivity" are hypothesis generation, not confirmation.

Self-check

1. In your EHR cohort, both preoperative opioid use and frailty increase the number of preoperative visits. Frailty also raises the risk of complications after THA. What bias could arise if you adjust for the visit count?
2. Blinded outcome assessment is impossible in an implant-comparison RCT (different implants look different on X-ray). What can you still blind?
3. You want the total effect of surgical approach on readmission. What happens if you adjust for length of stay on that causal pathway?

Answers

1. Collider bias. The visit count is a common effect of opioid use and frailty. Adjusting for it opens the path opioid use → visits ← frailty → complications. See Weiskopf et al. [8].

2. Blinded interpretation: the analyst pre-registers what conclusion each result pattern would support, then runs the analysis with group labels masked. Trial

reporting can also be blinded to primary-vs-secondary allocation until analysis is locked. See Järvinen et al. [7].

3. Adjustment blocks the pathway through length of stay. The estimate can move toward or away from the null. To interpret it as a controlled direct effect, specify the length of stay you would hold fixed and justify the assumptions, including adequate control of mediator–outcome confounding.

Anchor references [7,9,8]

Module 04 · Outcomes and measurement

Learning objectives. After this module you can choose primary and secondary outcomes that reflect the decision at stake; distinguish patient-important outcomes from surrogates; interpret MCID and PASS correctly.

Concept — brief. Outcomes are the currency of your paper. Choose them by what a patient and a clinician would change their decision over — not by what your database exports. Patient-reported outcome measures (PROMs) dominate arthroplasty because the operation is elective and the goal is function and pain relief, not mortality [10]. Core outcome sets (COS) — standardized minimum outcome bundles for a condition — reduce reporting heterogeneity and let readers compare studies [11,12].

Two thresholds are commonly confused:

- **MCID (minimal clinically important difference)** — the smallest change a patient perceives as meaningful.
- **PASS (patient acceptable symptom state)** — the level of function or symptom below which patients call themselves "well."

They measure different things. MCID is a *change* threshold; PASS is a *state* threshold. Reporting only mean change without either leaves the reader blind to whether your effect matters.

Checklist — outcomes

- ☐ One prespecified primary outcome, with a defined measurement instrument, timing, and analytic method.
- ☐ Secondary outcomes are prespecified and limited in number (headline of "we measured 27 things" is not a paper — it is a search).
- ☐ For PROMs, report distribution + proportion reaching MCID *and* PASS where thresholds exist.
- ☐ Adverse events use a standard nomenclature (Clavien-Dindo for surgical, CTCAE for oncologic, MedDRA for pharma).
- ☐ Consider registry- or COS-endorsed instruments to enable pooling.

- ☐ For time-to-event outcomes, prespecify censoring rules and how competing risks are handled.

Worked example — arthroplasty.

Question: *Does a new dual-mobility cup reduce dislocation at 2 years?*

Bad outcome: "any hip complication." Aggregates rare and non-comparable events; effect estimate uninterpretable.

Better: primary = incidence of dislocation requiring closed or open reduction within 2 years, per registry definition. Secondary = HOOS-JR change from baseline at 1 year (reported as mean change, %MCID, %PASS); revision for any cause at 2 years; instability-related readmission.

Pitfalls

- **Surrogate outcomes** (radiographic wear, biomarker level) can move independently of what the patient feels.
- **PROMs at fixed timepoints only** miss trajectory. If feasible, include one intermediate timepoint.
- **Composite endpoints** dilute effects when components move in opposite directions.
- **Ignoring the ceiling effect** in older, higher-functioning PROMs (e.g., WOMAC) can make a good implant look no better than adequate.

Self-check

1. What is the difference between MCID and PASS, and which one is a threshold for 'is this patient satisfied'?
2. Your composite endpoint is 'any complication.' Why is this a red flag?
3. Why does using a core outcome set (COS) increase the value of a small single-center study?

Answers

1. MCID (minimum clinically important difference) is the smallest change a patient perceives as meaningful; PASS (patient acceptable symptom state) is the outcome score above which the patient considers the symptom state acceptable. PASS is the 'satisfied' threshold. See Ogura et al. [13] and Katz et al. [14].

2. Composites hide effect size and treat rare-serious and common-minor events as equivalent. A 20% relative reduction in 'any complication' can be entirely driven by fewer superficial infections while deep infection rate is unchanged — which is the outcome the reader actually cares about.

3. Because a COS lets the study be pooled with others in a future meta-analysis. Idiosyncratic outcome choices strand data and preclude synthesis. See Kirkham et al. and the COMET initiative [11,12].

Anchor references [10,11,12]

Module 05 · Sample size and power

Learning objectives. After this module you can plan sample size from the anticipated effect and its clinical relevance; explain the difference between power for hypothesis testing and precision for estimation; use pilots correctly and stop overusing them.

Concept — brief. Sample size is a design choice, not a formula to satisfy a reviewer. Two philosophies:

- **Power-based** — pick a smallest effect worth detecting, an alpha and beta, and compute the N that gives adequate power. Traditional; still dominant for RCTs [13].
- **Precision-based** — pick the confidence interval width you can tolerate for the primary estimate; compute the N that yields that CI. More honest when the question is "how big is the effect" rather than "is the effect non-zero." Consistent with the estimation-first stance of Module 07.

Pilots exist to test feasibility — recruitment rate, procedure completion, protocol adherence — not to estimate the treatment effect. A pilot's effect-size estimate is too imprecise to plug into a downstream sample-size formula; using it as such produces underpowered main studies [14].

Conditional power — the probability of trial success given interim data — belongs to interim analyses in adaptive designs and is a separate topic [15].

Checklist — sample size

- ☐ State the **smallest effect worth detecting**, tied to the clinical decision (e.g., MCID for PROM difference).
- ☐ State the source of the assumed baseline event rate or SD, with a citation.
- ☐ State alpha, power, allocation ratio, one- vs. two-sided test.
- ☐ Account for expected loss to follow-up and non-adherence.
- ☐ For cluster designs, inflate by the design effect (ICC-based).
- ☐ For observational studies, plan for the events-per-variable rule of thumb in regression models (≥10-20 events per candidate covariate).
- ☐ If precision-based, state the target CI half-width and its rationale.
- ☐ Pilots: state their **feasibility** question, not their effect estimate.

Worked example — arthroplasty.

Question: *Does a new liner reduce revision at 5 years vs. standard?*

Baseline 5-year revision rate (registry): ~2%. Target relative reduction: 50% (i.e., to 1%) considered the smallest clinically meaningful. Two-sided α = 0.05, 80% power, 1:1 allocation. Required N per arm ≈ 2,300 (chi-square approximation). Loss to follow-up 10% → inflate to ~2,600 per arm. Feasibility test: no single-center RCT reaches this. Reframe: registry cohort with TTE, precision-based, target 95% CI half-width for hazard ratio ≤ 0.3 on the log scale.

Pitfalls

- **Power calculated from the observed effect** post hoc is meaningless.
- **Assumed effect sizes borrowed from prior RCTs** are often optimistic (regression to the mean).
- **Ignoring non-adherence** overstates achievable power.
- **Multiple primary outcomes** without alpha adjustment or hierarchical testing inflate type I error.

Self-check

1. Your sample size calculation gives N=200 for a two-arm superiority trial at 80% power. You expect 15% dropout. What N do you enroll?
2. You have 82 outcome events and want to fit a logistic-regression model. How many candidate covariates can you consider?
3. A trial reports 'no significant difference (p=0.11).' What must you look at to interpret this correctly?

Answers

1. Roughly N/(1-dropout) per arm — 200/0.85 = 236, round up to 240 per arm. Do not just enroll 200 total and lose power; do not enroll 240 total and split evenly (allocate 240 per arm).

2. About 8 (using the 10 events per variable rule of thumb; more recent work suggests 20 EPV for better calibration). Fitting a model with 15 predictors on 82 events overfits and yields optimistic AUC. See Riley et al. and Van Smeden et al. [42, 45].

3. The 95% CI on the effect estimate, the a priori MCID, and the powered effect size. 'No significant difference' with a wide CI spanning clinically important effects is uninformative, not negative. See ASA statement on p-values [22].

Anchor references [13,14,15]

Module 06 · Data management

Learning objectives. After this module you can build a data pipeline that another investigator could take over next week; distinguish data collection, cleaning, and provenance; use standard tooling instead of Excel-and-hope.

Concept — brief. Data management is where reproducibility either happens or it does not. Three layers:

1. **Capture** — the tool that ingests observations (REDCap, custom EDC, registry export) [16,17].
2. **Cleaning** — the deterministic transformation from raw to analyzable, documented as code.
3. **Provenance** — the audit trail that lets you say what happened to any value at any point.

The single highest-yield discipline: **cleaning is code, not manual editing.** Every deletion, recode, and merge is a scripted step. The raw file is immutable; the analytic file is regenerable [18].

Checklist — data management

- ☐ Data dictionary written before collection (variable names, types, allowed values, source instrument).
- ☐ Capture platform validated (skip logic, range checks, required fields, audit trail).
- ☐ Raw exports stored read-only with datestamped filenames.
- ☐ Cleaning implemented as a script (R, Python, Stata, SAS — pick one). Manual edits are forbidden.
- ☐ Every derived variable's formula is in the script, with a comment naming the source variables.
- ☐ Version control (git) for the code; DVC or equivalent for data snapshots if datasets are large.
- ☐ Missingness pattern examined before analysis; missing-data mechanism (MCAR/MAR/MNAR) assessed.
- ☐ Data-sharing plan (repository, access control, DUA) decided at protocol stage, not after acceptance.

Worked example — arthroplasty.

You are running a multicenter cohort of revision THAs. Each site enters into REDCap. A weekly export is downloaded, timestamped, and stored read-only. A single R script:

1. Reads raw REDCap export.
2. Applies range checks (e.g., stem size 6–24 mm; anything outside → flagged for source review, not silently deleted).
3. Recodes implant IDs to a canonical registry code from a lookup table (in git).

4. Derives per-patient composite features (Charlson from ICD codes; time-to-revision from index date).
5. Writes an analytic file with a version tag matching the git commit hash.

Any collaborator can rerun the pipeline and reproduce the analytic file bit-for-bit. Any anomaly can be traced back to its source row and the transformation that touched it.

Pitfalls

- **Excel drift** — sorting a subset of columns silently misaligns rows. It has happened to everyone. Cost your team a paper.
- **"Cleaned" file with no script** — you cannot reproduce it, and neither can anyone else.
- **Free-text fields** used for structured concepts. Enforce controlled vocabularies at capture.
- **Silent missing data recoding** (e.g., blank → 0). Missing is not zero. Handle explicitly.

Self-check

1. You are building a REDCap database for a multi-center registry. Name the three fields you must design before you enter a single patient.
2. Why is 'missing = 0' a defect, not a convenience?
3. You are asked to share the dataset with a collaborator overseas. What are two things you check before you export?

Answers

1. First: A stable, non-repeating patient identifier that is not the MRN. Second: An event timestamp with time zone. Third: A source-tag field naming which site and which extraction date the row came from. Everything else can be added later; these three cannot be retrofitted cleanly.

2. Because it silently converts a data-quality problem into an analytic one. Downstream summaries treat imputed zeros as measured zeros and bias every derived statistic. Use explicit missing indicators and handle missingness in the analysis (Module 07). See Austin et al. [21].

3. First: Data-use agreement or ethics amendment permitting the transfer. Second: De-identification level appropriate to the recipient jurisdiction (HIPAA Safe Harbor, GDPR pseudonymization, or a signed DUA for full identifiers). Log what you sent, to whom, and when. See Pellen et al. [18].

Anchor references [16,17,18]

Module 07 · Analysis and estimation

Learning objectives. After this module you can present an effect estimate with a confidence interval and let the reader do the interpretive work; recognize why the p-value has been widely misused; handle missing data with multiple imputation instead of complete-case shortcuts.

Concept — brief. The single most consequential culture change in clinical research over the last decade is the shift from significance testing to **estimation**. Report the magnitude of the effect and its precision — the effect size and the confidence interval — as the primary result. The p-value is a secondary quantity, if reported at all [19,20,21].

This is the discipline behind OrthoScienus's *Estimation First* layer. A $p < 0.05$ without an effect size is uninterpretable. A $p = 0.08$ with a large effect and wide CI is also uninterpretable — but for the opposite reason.

Bayesian analysis frames the same question as a posterior probability distribution over the effect. In orthopaedic decision-making, where prior evidence is often strong (registries, prior RCTs), Bayesian methods are increasingly viable and increasingly published in general ortho journals [22].

Missing data is not an inconvenience — it is a data feature. Under MAR, multiple imputation is the standard defensible approach; complete-case analysis silently selects your sample [23].

Checklist — analysis and estimation

- ☐ Primary result is an **effect size with 95% CI**, not a p-value.
- ☐ Effect size units are meaningful to a clinician (absolute difference, hazard ratio, MCID multiples — not "standardized mean difference" alone).
- ☐ If p-values are reported, they are secondary and never dichotomized as "significant / not significant."
- ☐ Missing data mechanism assessed and named. Multiple imputation used under MAR; sensitivity analysis for MNAR.
- ☐ Model assumptions stated and checked (linearity, proportional hazards, calibration).
- ☐ Analysis plan pre-specified (Module 12); deviations flagged and justified.
- ☐ Consider a Bayesian analysis with a weakly informative prior as a sensitivity analysis when a strong prior exists.

Worked example — arthroplasty.

You compare 1-year HOOS-JR change between two approaches in a registry cohort.

- Bad: "The anterior group improved more ($p = 0.03$)."

- Better: “Adjusted mean difference in HOOS-JR at 1 year was 3.1 points favoring anterior (95% CI 0.4 to 5.8), below the MCID of ~9 points. We do not interpret this as a clinically meaningful advantage.”

The second version is honest about a statistically detectable but clinically small effect. The first version misleads.

Pitfalls

- **P < 0.05 as a proxy for truth.** It never was.
- **“Trend toward significance”** — there is no such thing. Report the estimate and CI.
- **Adjusting until significant** — garden of forking paths. Prespecify.
- **Complete-case analysis without justification** — silent selection bias.
- **Reporting only relative effects** (RR, OR) without absolute numbers — hides clinical magnitude.

Self-check

1. Why is ‘p < 0.05’ a poor primary output for a clinical study?
2. Multiple imputation vs. complete-case analysis — when is complete-case actually valid?
3. A subgroup analysis shows the treatment effect is significant in men (p=0.03) but not women (p=0.20). What are you allowed to conclude?

Answers

1. Because it collapses effect size and precision into a binary that is almost always misinterpreted. Report the effect estimate with a 95% CI as the primary result; the p-value is a secondary summary of that estimate against a null. See Wasserstein & Lazar (ASA) [22].

2. When data are missing completely at random (MCAR) and the missingness is small. If missingness relates to exposure or outcome (MAR or MNAR), complete-case biases the estimate. Multiple imputation is more general but not a cure for MNAR. See Austin et al. [21].

3. Almost nothing about subgroup-specific effects without a prespecified interaction test. A significant subgroup and a non-significant subgroup can be entirely compatible with a homogeneous effect that is just noisier in the smaller subgroup. Test interaction; do not compare within-subgroup p-values.

Anchor references [19,20,21,22,23]

Module 08 · Causal inference from observational data

Learning objectives. After this module you can draw a directed acyclic graph (DAG) to identify confounders and colliders; state the assumptions that justify a causal claim from observational data; recognize when the target-trial framework applies to your dataset.

Concept — brief. Observational data cannot deliver causal answers by adjustment alone. It can when three things hold: (i) the correct set of confounders is measured and adjusted; (ii) there is no unmeasured confounding of consequence; (iii) the counterfactual contrast is well-defined [24].

A **DAG** is the working tool for (i). Nodes are variables; arrows are direct causal effects. The minimum sufficient adjustment set is the set that blocks all backdoor paths without opening any new ones (via colliders). DAGs are the difference between "I adjusted for everything I had" and "I adjusted for the variables that identify the causal effect" [25,26].

Target trial emulation was introduced in Module 02 as a design-selection heuristic. Here it becomes the analysis discipline that keeps observational effect estimates honest: use inverse probability weighting or matching to approximate randomization, align time zero with eligibility and treatment assignment, and report per-protocol and intention-to-treat analogues in parallel so the reader sees both the biological effect and the pragmatic one [4,5,27].

Checklist — causal inference

- ☐ DAG drawn before analysis, sharing among coauthors and reviewers.
- ☐ Minimum sufficient adjustment set identified from the DAG; not "the kitchen sink."
- ☐ Colliders identified and *excluded* from adjustment.
- ☐ Identify mediators and leave them out of the adjustment set when estimating the total effect. For a controlled direct effect, specify the mediator level you would hold fixed and justify the assumptions.
- ☐ Time zero, eligibility, and treatment assignment are aligned (target trial).
- ☐ Positivity checked — every combination of covariates has both treated and untreated patients.
- ☐ Unmeasured confounding addressed via E-value or negative control analysis.
- ☐ Sensitivity analyses varying the confounder set are prespecified.

Worked example — arthroplasty.

Question: *Does chronic opioid use before THA increase 90-day readmission?*

Naïve model: readmission ~ opioid + age + sex + BMI + ASA. Reviewer asks: "Did you adjust for pain scores?" Draw the DAG.

- Chronic pain → opioid use.
- Chronic pain → readmission (via functional decline, complications).

- Opioid use → readmission (via respiratory events, delirium).
- A baseline patient-reported function score measured before opioid use began is a **proxy for chronic pain**.

Adjusting for the function score may reduce confounding, but patients with the same score can still differ in chronic pain. A score collected after opioid use began may itself reflect treatment. Leave *postoperative pain* out of the total-effect model if opioid use changes pain and pain affects readmission. Suppose opioid use and frailty both influence *discharge disposition*, and frailty also increases readmission risk. Disposition is a collider on the path opioid use → disposition ← frailty → readmission.

Pitfalls

- **"I controlled for confounders"** without a DAG is a claim, not evidence.
- **Automated variable selection** (stepwise, LASSO by default) is not causal identification.
- **Propensity scores** address measured confounding — no more.
- **Positivity violations** yield extrapolation beyond your data; check overlap plots.

Self-check

1. In a study of sustained opioid use, pain after the first prescription affects later prescribing and readmission. The first prescription also affects pain. How would you estimate the total effect of sustained use?
2. What does an E-value of 1.8 tell you about your effect estimate?
3. You match 400 exposed to 400 unexposed on propensity score and report a treatment effect. What have you not addressed?

Answers

1. Pain can mediate the effect of early treatment and confound the effect of later treatment. Simply adjusting for it in a regression can block part of the early effect. Define the treatment strategies and consider a longitudinal method such as the g-formula if its assumptions are plausible. Randomizing the first prescription addresses baseline confounding. Its effect can still operate through later pain.

2. That an unmeasured confounder would need to be associated with both exposure and outcome by a risk ratio of at least 1.8 (above and beyond the measured confounders) to fully explain away your observed effect. Small E-values mean the result is fragile to plausible unmeasured confounding; large E-values mean it is robust.

3. Unmeasured confounding. Propensity-score matching balances measured covariates only. Also positivity — check that both treated and untreated patients exist across the covariate space (overlap plots). See Inoue et al. [9].

Anchor references [24,25,26,27]

Module 09 · Systematic review workflow

Learning objectives. After this module you can run a systematic review that a peer could reproduce, and know when a scoping or living review is the honest form.

Concept — brief. A systematic review is a study, not a literature review. It has a prespecified question (PICO), an explicit search, transparent screening, standardized extraction, and a formal synthesis — with or without meta-analysis. **PRISMA 2020** is the reporting standard for systematic reviews of interventions [28,29]. Network meta-analyses have their own extension, PRISMA-NMA [30]. Living systematic reviews update continuously; they are the correct form when evidence is accumulating fast (COVID, cell therapy, PROMIS instrumentation).

Search strategy is where most reviews fail silently. Single-database (usually PubMed) searches miss 20–40% of eligible studies. Optimal database combinations, especially Embase + MEDLINE, materially reduce this [31].

Checklist — systematic review

- ☐ Protocol registered on PROSPERO before screening.
- ☐ PICO documented; eligibility criteria unambiguous.
- ☐ Search strategy peer-reviewed (PRESS checklist), covers ≥2 databases (typically MEDLINE + Embase + trial registries).
- ☐ Two independent screeners at title/abstract and full-text stages; disagreements adjudicated.
- ☐ Extraction form piloted on 5–10 studies; agreement checked.
- ☐ Risk of bias assessed with an appropriate tool (Module 10).
- ☐ Certainty of evidence rated with GRADE (Module 10).
- ☐ PRISMA 2020 flow diagram and checklist included.
- ☐ Deviations from protocol documented and justified.

Worked example — arthroplasty.

Question: *In patients undergoing revision TKA for infection, does one-stage vs. two-stage exchange result in higher infection-free survival at 5 years?*

- Protocol: PROSPERO registered; PICO defined; primary outcome infection-free survival at 5 years.
- Search: MEDLINE + Embase + CENTRAL + ClinicalTrials.gov. Native database syntax, not a copy of a Google query. Peer-reviewed with an information specialist.
- Screening: two reviewers, blinded to each other; Covidence or Rayyan for workflow.
- Extraction: piloted on the first five included studies; refined; then applied to all.

- Risk of bias: ROBINS-I (most studies observational); Cochrane RoB 2 for the rare RCT.
- Synthesis: random-effects meta-analysis of Kaplan-Meier–derived hazard ratios if feasible; narrative synthesis otherwise. Heterogeneity via I^2 and prediction interval.
- Certainty: GRADE-rated per outcome.
- Report: PRISMA 2020.

Pitfalls

- **PubMed-only search** — misses ~20–40% of eligible studies.
- **Post-hoc PICO changes** during screening — you are conducting a scoping review at that point; call it that.
- **Vote counting** — counting "positive" vs. "negative" studies is not synthesis.
- **Publication bias** ignored — funnel plot + Egger's test at minimum for ≥10 studies; contour-enhanced funnel plot preferred.
- **Overlap in umbrella reviews** — count each primary study once.

Self-check

1. You screen 1,200 titles and abstracts and include 34 studies. Two reviewers disagree on 60 records. What do you report?
2. When is a scoping review the honest form, not a systematic review?
3. Why is a living systematic review a good match for a fast-moving evidence area but a bad match for a stable one?

Answers

1. The number of disagreements, the resolution mechanism (third reviewer, discussion), and the inter-rater agreement statistic (Cohen's kappa or percent agreement) at title/abstract and full-text stages. PRISMA 2020 requires it. See Page et al. [28,29].

2. When the question is a mapping question ('what has been studied, how, in what populations') rather than an effect question. Scoping reviews describe the literature; systematic reviews answer a defined question and typically synthesize effects. Do not run a scoping review and then claim causal conclusions from it.

3. Because the maintenance cost is only justified when new evidence arrives frequently enough to change conclusions. On a stable topic, a living review burns effort updating things that will not change and gives readers a false sense of active vigilance.

Anchor references [28,29,30,31]

Module 10 · Risk of bias and certainty of evidence

Learning objectives. After this module you can apply the correct risk-of-bias tool to each study design in your review; rate the certainty of a body of evidence with GRADE; explain what a "moderate certainty" rating means to a clinician.

Concept — brief. Two levels of appraisal:

- **Study level:** risk of bias per study. RCTs → **RoB 2** [32]. Non-randomized studies of interventions → ROBINS-I. Cohorts of exposure → ROBINS-E. Diagnostic accuracy → QUADAS-2. Prognostic models → PROBAST.
- **Outcome level:** certainty of evidence across studies for each outcome. **GRADE** rates certainty as High, Moderate, Low, or Very Low based on risk of bias, inconsistency, indirectness, imprecision, and publication bias. In 2025, the GRADE Working Group released a refreshed **Core GRADE** approach that formalized rating principles [33]. **CINeMA** applies analogous ratings to network meta-analyses [34].

"Moderate certainty" means: further research is likely to have an important impact on the estimate. It is not a synonym for "moderate effect."

Checklist — RoB and certainty

- ☐ Correct tool per study design (RoB 2 for RCT; ROBINS-I for non-randomized).
- ☐ Two independent raters; disagreements adjudicated.
- ☐ Domain-level judgments, then overall — not a checklist total.
- ☐ GRADE Summary of Findings table per outcome.
- ☐ Each GRADE domain explicitly rated (RoB, inconsistency, indirectness, imprecision, publication bias); rationale documented.
- ☐ For network meta-analyses, use CINeMA in place of standard GRADE.
- ☐ Communicate certainty in plain language ("we are moderately confident that…").

Worked example — arthroplasty.

A meta-analysis of 6 RCTs and 12 cohort studies of tranexamic acid in TKA.

- Per-study RoB: RoB 2 for RCTs, ROBINS-I for cohorts.
- Meta-analysis stratified by design.
- GRADE for the primary outcome (transfusion): start High for RCTs → downgrade for imprecision if CI crosses no effect and MCID; downgrade for inconsistency if I^2 > 60% with clinical heterogeneity; downgrade for publication bias if funnel asymmetry.
- Summary of Findings table: absolute effect (per 1000), relative effect, N studies, N patients, certainty.

Pitfalls

- **Wrong tool** — Newcastle-Ottawa Scale for observational studies is no longer state of the art; use ROBINS-I.
- **Cochrane RoB 1** — retired; use RoB 2.
- **Downgrading for a single small study** — inconsistency needs at least a few studies to judge.
- **Blanket "Low" ratings** — a GRADE table with everything Low signals over-severity or an under-designed review.
- **Not reporting Summary of Findings** — GRADE without a SoF is halfway.

Self-check

1. Your RoB 2 assessment gives 'some concerns' for allocation concealment in 4 of 6 trials. What GRADE domain does this affect and by how much?
2. Distinguish 'imprecision' from 'inconsistency' in GRADE.
3. What is the one thing a GRADE Summary of Findings table forces you to declare that a narrative review usually hides?

Answers

1. Risk of bias domain in GRADE. Serious risk-of-bias concerns typically warrant a one-level downgrade (from High to Moderate); the exact call depends on how many trials contribute to the pooled estimate and whether the concern is likely to bias the effect. Document the rationale. See Sterne et al. [32].

2. Imprecision is about the width of the CI on the pooled estimate (too wide to be actionable). Inconsistency is about heterogeneity across studies (I^2, prediction interval). They can co-occur but are separate downgrade reasons and should be documented independently.

3. Absolute effects per 1,000 patients (or per unit), by outcome, tied to a stated certainty rating. It converts vague 'evidence suggests' into a number the reader can act on. See Guyatt et al. [34].

Anchor references [32,33,34]

Module 11 · Reporting guidelines

Learning objectives. After this module you can pick the correct EQUATOR reporting guideline for your study before you write; use it as a scaffold for the manuscript, not a submission checklist.

Concept — brief. Reporting guidelines exist to make studies reviewable. Use them at the start — as an outline for the paper — not at the end as a compliance exercise.

Core, current standards:

Study design	Reporting guideline	Current version	Refs
Randomized trial	CONSORT	2025	[35]
Trial protocol	SPIRIT	2025	[36]
Observational (cohort/case-control/cross-sectional)	STROBE	2007 (refresh underway)	—
Systematic review	PRISMA	2020	[28,29]
Network meta-analysis	PRISMA-NMA	2015	[30]
Prediction model (regression or ML)	TRIPOD+AI	2024	[37]
AI-enabled trial	CONSORT-AI	2020	[38]
Diagnostic accuracy	STARD	2015	—
Case report	CARE	2013	—
Clinical practice guideline	AGREE II / RIGHT	2017 / 2017	—
Qualitative research	SRQR / COREQ	2014 / 2007	—
Preclinical animal	ARRIVE 2.0	2020	—

The EQUATOR Network catalogs the current versions and extensions; check before writing, not after.

Checklist — reporting

- ☐ Guideline identified before writing.
- ☐ Manuscript sections mirror the checklist order.
- ☐ Every checklist item addressed in the paper or explicitly noted as N/A.
- ☐ Flow diagram / figure included where required (CONSORT flow, PRISMA flow, STARD flow).
- ☐ Reporting checklist uploaded as a supplement at submission.
- ☐ For AI or model studies, TRIPOD+AI or CONSORT-AI extensions used from the outset.

Worked example — arthroplasty.

A single-center retrospective cohort of dual-mobility THA in patients with a history of instability.

- Guideline: STROBE (observational cohort).
- Flow: eligible patients → excluded → included → analyzed.
- Item-by-item mapping to STROBE at first draft.

- Add TRIPOD+AI checklist if any prediction model is fit (e.g., for revision risk).
- Reporting checklist uploaded at submission.

Pitfalls

- **Guideline as post-hoc paperwork.** It is a writing scaffold; if you find yourself unable to fill an item, you have a study design gap, not a reporting gap.
- **STROBE ≠ CONSORT.** Pick per design.
- **Outdated version** (CONSORT 2010 when 2025 exists) — signals inattention.
- **Ignoring extensions** (CONSORT-AI, PRISMA-NMA) — signals unfamiliarity with your own methodology.

Self-check

1. You are writing a retrospective cohort of dual-mobility THA. Which reporting guideline governs it, and which extension do you also check?
2. A journal asks you to submit a 'CONSORT-AI checklist' with your ML-prediction model paper. Should you comply?
3. You cannot complete a STROBE checklist item because the data are unavailable. What do you do?

Answers

1. STROBE governs the manuscript. Check STROBE-RECORD if the data source is a routinely collected health database (registry, EHR), because it adds items on database provenance and record linkage that STROBE alone under-specifies.

2. Only if your paper is actually a trial of an AI-enabled intervention. If it is a development-and-validation paper without a trial, use TRIPOD+AI [37] instead. CONSORT-AI is a CONSORT extension for trials; using it for non-trial work signals unfamiliarity.

3. Mark N/A with an explicit brief justification in the checklist (e.g., 'exposure ascertainment timing not recorded in source registry'). Do not leave the item blank and do not fabricate. Reviewers and readers use the checklist to see design gaps; hiding them is dishonest.

Anchor references [28,29,35,36,37,38]

Module 12 · Reproducibility and workflow

Learning objectives. After this module you can register a protocol before enrolling; share code and, where possible, data; build a workflow that a stranger could reproduce a year from now.

Concept — brief. Reproducibility is the outer skin of the whole course. It comprises:

- **Prospective registration** of protocol and analysis plan (ClinicalTrials.gov, OSF, PROSPERO).
- **Version-controlled code** for cleaning, analysis, and reporting.
- **Data sharing** where governance permits, with a persistent identifier.
- **Preregistered deviations** — when you must change the plan, say so and why.

The reproducibility crisis is neither hyperbole nor solved. In clinical and translational research it is largely a statistics-review problem (Module 07's misused p-values) compounded by opaque workflows [39]. Open, reproducible epidemiology is technically feasible today; the barrier is habit, not tooling [40,41].

Checklist — reproducibility

- ☐ Protocol registered before enrollment / data extraction.
- ☐ Statistical analysis plan finalized before database lock or unblinding.
- ☐ Code under version control (git); repository linked at publication.
- ☐ Data sharing: consented and DUA-covered participant data deposited in a repository (Dryad, Zenodo, institutional) or, where restricted, metadata + access mechanism disclosed.
- ☐ Environment captured (renv, conda, container) so results can be regenerated.
- ☐ Preregistered analytic plan cited in the methods; deviations transparently reported.
- ☐ For registries, honor the registry's data-sharing rules.

Worked example — arthroplasty.

A multicenter cohort of revision THA for instability.

- Protocol on OSF with a hash of the analytic plan; DOI in the methods.
- Cleaning + analysis + figure code in a git repository; commit hash referenced in the paper.
- De-identified analytic dataset on Zenodo where consent allows; otherwise, a synthetic dataset that reproduces the analysis structure with the code deposited.
- Deviations from the pre-registered analysis listed in a supplement with rationale.
- Container (Docker) with the R/Python environment specified for regeneration.

Pitfalls

- **"Available on request"** — not sharing.
- **Registered post-hoc** — not registration.
- **Analytic file with no code** — not reproducible.
- **Sharing without a license** — legally ambiguous.
- **Ignoring institutional data governance** — the fastest path to a retraction.

Self-check

1. You register a trial protocol on ClinicalTrials.gov but change the primary outcome after enrollment starts. What is your obligation?
2. What is the minimum reproducibility artifact you should ship with an observational study?
3. Why is a Git-tracked analysis directory not, by itself, reproducibility?

Answers

1. Update the registration record with the change and a dated rationale; disclose the change and its rationale in the published manuscript. Silent primary-outcome switching is a documented cause of publication bias and reader distrust.

2. The final analytic dataset schema, the analysis code (with a fixed random seed), the software environment (lockfile or container), and a run script that reproduces the reported figures and tables from the raw inputs. Data itself may be restricted; the code and specification should not be.

3. Because Git tracks code, not data provenance or environment. Add data versioning (DVC or similar), an environment lockfile, and a documented run entry point. Reproducibility is the union of code + data + environment + instructions, not any one of them.

Anchor references [39,40,41]

Module 13 · Scientific writing

Learning objectives. After this module you can write an IMRaD manuscript that a busy reader can navigate; write an abstract that survives skim-reading; write a discussion that argues from your data, not around it.

Concept — brief. IMRaD (Introduction, Methods, Results, and Discussion) is not a template — it is a discipline [42,43]. Each section has a specific job:

- **Introduction:** why the question, in three paragraphs. What is known → what is missing → what this paper does.
- **Methods:** what you did in enough detail that a peer could replicate. Written first, in past tense.
- **Results:** what you found, without interpretation. Numbers, effect sizes, CIs. Tables and figures do the heavy lifting.
- **Discussion:** what your findings mean, what the limitations are, what the reader should do next. Argue from your data.

The abstract is not a summary — it is the paper's front door. Structure it (Background / Methods / Results / Conclusion) with numbers, not adjectives.

Checklist — scientific writing

- ☐ Methods written first, from the protocol.
- ☐ Results contain no interpretation; every number in the text is in a table or figure.
- ☐ Effect sizes with CIs are the primary results; p-values are not the headline.
- ☐ Discussion opens with a one-sentence summary of the main finding, not "This is the first study to…"
- ☐ Limitations section names specific threats to validity and how they were addressed.
- ☐ Abstract mirrors the paper's structure; conclusions match results.
- ☐ Reporting guideline (Module 11) followed section by section.

Worked example — arthroplasty.

Discussion opening — two versions of the same finding:

- Puffed: "This landmark study is the first to demonstrate that the novel technique offers substantial benefits…"
- Direct: "In this multicenter cohort of 4,182 primary THAs, dual-mobility cups were associated with a 40% lower 2-year dislocation rate compared with fixed-bearing cups (adjusted HR 0.60, 95% CI 0.44 to 0.82). The absolute risk difference was 1.2 percentage points, and the effect was consistent across age, sex, and approach subgroups."

The second version is shorter, states the effect, and lets the reader judge.

Pitfalls

- **Introduction that reviews the field for two pages** — you are writing a paper, not a textbook chapter.
- **Results with interpretation** ("we found a strikingly large effect") — save for discussion.
- **Discussion that ignores contradictory literature** — reviewers notice.
- **Limitations that are humble-brags** ("the only limitation is our large sample size") — insulting.
- **AI-written abstracts** — increasingly detectable, and increasingly restricted by journal policy [44] (see Module 17).

Self-check

1. Your Discussion opens: 'In our large series of 84 hips…' What is wrong?
2. You are writing the Methods section and cannot remember whether you used mean or median. What do you do?
3. A reviewer says 'the Introduction is too long.' What is the usual real problem?

Answers

1. ‘Large’ is a size claim that 84 hips does not support in most arthroplasty contexts. Delete the adjective. Report N and let the reader judge. Adjectival hedging or inflation is the first place a suspicious reviewer starts.

2. Go back to the analysis code or the dataset and check. Never estimate or infer a Methods detail from memory. Then, in the code, add a comment noting the choice so the next person (including future you) does not have to re-derive it.

3. The Introduction is reprising the literature review instead of setting up the question. Tighten to: what is known, what is missing, what this study adds. Three paragraphs, not eight. See Willis [43].

Anchor references [42,43]

Module 14 · Data visualization

Learning objectives. After this module you can pick the right chart for the right claim; produce a forest plot, calibration plot, and decision-curve analysis that read cleanly; avoid the visualization anti-patterns that hide effect sizes.

Concept — brief. A good figure lets the reader see the effect in one glance. A bad one hides it under color and chartjunk. Three high-yield charts in clinical research:

- **Forest plot** — the standard for meta-analyses and subgroup analyses. Effect size + 95% CI per study; overall estimate with prediction interval.
- **Calibration plot** — the honest check on a prediction model: predicted probability vs. observed rate, ideally with a smoothed calibration curve and confidence band [45] (Module 18).
- **Decision curve analysis** — plots net benefit across a range of threshold probabilities; answers “is this model or test worth using in decisions?” — a question ROC AUC does not answer [46,47].

Chart-type selection

Claim you are making	First-choice chart	Second-choice	Anti-pattern to avoid
Effect size + precision across studies	Forest plot	Table with CI	Bar chart of point estimates
Model discrimination	ROC curve (with AUC)	Precision-recall (rare outcomes)	AUC as a lone number
Model calibration	Smoothed	Reliability diagram	Hosmer-Lemeshow

Claim you are making	First-choice chart	Second-choice	Anti-pattern to avoid
	calibration plot + histogram		only
Clinical utility of a model	Decision curve analysis	Net reclassification (with caveats)	Sensitivity/specificity table alone
Time-to-event	Kaplan-Meier + at-risk table	Cumulative incidence (if competing risks)	KM without at-risk numbers
Distribution comparison across groups	Box plot with jitter, or violin	Overlaid histograms	Bar chart with SEM
Change in continuous outcome across time	Line plot with CI ribbon	Small multiples per subgroup	Bar chart with error bars
Multiple subgroup effects	Small multiples or forest plot	Table	One crowded panel
Proportions across a few categories	Bar chart with error bars	Stacked bar for parts of whole	Pie chart
Continuous vs. continuous	Scatter + smoothed regression + CI	Hex-bin (large N)	Line-of-best-fit alone

Checklist — data visualization

- ☐ One claim per figure.
- ☐ Effect size and precision (CI) visible.
- ☐ Axes labeled with units; scale chosen to show variability, not to compress it.
- ☐ Color used to encode a variable, not to decorate.
- ☐ Colorblind-safe palette (viridis, ColorBrewer).
- ☐ Data-to-ink ratio maximized (Tufte). Chartjunk minimized.
- ☐ Small multiples for subgroup comparisons, not one crowded panel.
- ☐ For predictive models: report and plot calibration alongside discrimination.
- ☐ For clinical utility: report and plot decision curves.

Worked example — arthroplasty.

A 5-year revision-risk model for THA has AUC 0.78 — sounds good. Add:

- Calibration plot: predicted probability vs. observed 5-year revision, deciles of risk. Reveals whether the model over- or under-predicts in the high-risk group where a decision would actually be made.

- Decision curve: net benefit at threshold probabilities 5%, 10%, 20%. Reveals whether the model beats “treat all” or “treat none” in the decision range.

If the decision curve shows no net benefit above threshold 5%, the AUC is fine and the model is clinically useless. That is the point of DCA.

Pitfalls

- **Bar charts with SEM** — hide distribution; use box plots, violin plots, or raw data with jitter.
- **Truncated y-axis** — inflates apparent effect.
- **Pie charts for >3 categories** — unreadable.
- **AUC alone** for a prediction model — hides miscalibration and lack of net benefit.
- **Kaplan-Meier without at-risk table** — precision at late timepoints is invisible.

Self-check

1. Your model has AUC 0.82 and a decision curve showing no net benefit above threshold probability 5%. What do you report?
2. Why is a bar chart with SEM error bars a poor summary of a continuous outcome distribution?
3. You have a Kaplan-Meier curve extending to 8 years. What single element do you include to protect the reader from over-reading the tail?

Answers

1. Both. AUC alone is misleading because it does not answer ‘is this useful in decisions?’ Report AUC for discrimination, calibration plot for reliability, and decision curve for clinical utility. If DCA shows no net benefit in the decision range, say so and question deployment. See Van Calster et al. and Vickers et al. [45, 46, 47].

2. Because it hides distribution shape and treats the SEM as if it were the uncertainty of interest (usually it is not — the SD or the 95% CI of the estimate is). Use box plots with jittered points, or violin plots, to show shape.

3. The at-risk table beneath the x-axis. Late-timepoint estimates from a KM curve depend on very few remaining patients; the at-risk numbers make that dependence visible.

Anchor references [46,47]

Module 15 · Peer review

Learning objectives. After this module you can review a manuscript in a way that improves it and does not waste the authors' time; respond to reviewer comments in a way that resolves rather than escalates; recognize a good peer review from your own drafts.

Concept — brief. Peer review is a professional obligation that most surgeon-scientists learn by mimicry. Two roles:

- **As reviewer:** identify what the paper claims, whether the methods can support that claim, whether the reporting is sufficient for replication, and whether the paper adds to the field [48,49].
- **As author:** treat each reviewer comment as either (a) a request to change the paper, (b) a request to add a citation, (c) a request to clarify wording, or (d) a disagreement. Address (a)-(c) in the manuscript and the response letter. For (d), push back respectfully with data or logic, not with tone.

Checklist — as reviewer

- ☐ Read the paper twice — once for the claim, once for the methods.
- ☐ Comment on the question, the design, the analysis, the interpretation, and the reporting.
- ☐ Distinguish essential comments (must be addressed) from minor comments (nice to have).
- ☐ Do not require citations to your own work unless directly relevant.
- ☐ Do not require additional studies beyond the paper's scope.
- ☐ Comment on statistical adequacy specifically (design, analysis, interpretation).
- ☐ Respect the timeline; decline promptly if you cannot review within it.
- ☐ Disclose COI (Module 17).

Checklist — as author responding to reviews

- ☐ Response letter mirrors reviewer comments verbatim, with author response inline.
- ☐ Every change is quoted with location (page/line).
- ☐ Disagreements are supported by data or references, not by tone.
- ☐ Thank the reviewer once; do not grovel.
- ☐ Submit tracked-changes manuscript alongside clean version.

Worked example 1 — responding to a scope-expansion demand.

Reviewer 2 writes: "The sample size is small and the authors should recruit more patients."

Bad response: "We agree the sample size is a limitation and have added this to the discussion."

Better response: “The study was powered a priori for the primary outcome at N=180 per arm (Methods, Section 2.4, page 6). We agree the CI on the secondary outcome is wide and we now report this explicitly. Additional recruitment is outside the scope of this paper; the trial is closed. We have added the CI width and its implications to the Discussion (page 12, lines 8-15).”

The second version pushes back on an inappropriate request while acknowledging the underlying concern.

Worked example 2 — reviewing an arthroplasty registry paper.

A manuscript reports “cementless stems have 1.6x higher revision risk than cemented in women >75.” The authors adjusted for age, BMI, ASA, and surgeon volume using a Cox model on 8-year registry data.

A useful review does not ask for more analyses. It asks the three questions that determine whether the claim survives:

- **Time zero and immortal-time bias.** Was exposure (stem type) defined at surgery, before any post-surgery events could influence the classification? If any patient was reclassified after the fact, hazards are biased downward for the exposed group.
- **Case-mix confounding not captured by the adjustment set.** Bone quality, comorbidity burden, and surgeon-selection preferences drive stem choice in this population; adjusting for age/BMI/ASA/volume leaves substantial residual confounding. Ask the authors to add a DAG (Module 08) and an E-value for the primary effect.
- **Competing risk of death.** In women >75 followed for 8 years, mortality is a competing risk for revision. A Cox model estimates cause-specific hazards; the clinically meaningful quantity is the cumulative incidence function. Ask for a Fine-Gray subdistribution model or CIF estimates alongside the Cox HR.

Essential comments: DAG + E-value; competing-risks re-analysis. Minor: report absolute revision risks per 1,000 patient-years, not only HRs. Not appropriate: “add a matched RCT for validation.”

Worked example 3 — handling a reviewer who requests self-citation.

Reviewer comment: “The authors should cite Smith et al. 2021, 2022, 2023, and 2024, which are the seminal papers in this area.”

All four are by the same author group, none are on the exact question at hand, and one is a conference abstract.

Response: “We appreciate the reviewer pointing us to this body of work. We have added Smith et al. 2023 [ref], which addresses [specific relevant aspect]. The 2021 and 2024 papers focus on [different population/exposure] and, while adjacent, are not directly on the comparator question here. The 2022 conference abstract is not indexed and is

superseded by the 2023 full paper we now cite. Should the editor prefer, we can add the additional references as supplementary reading."

This version does not stonewall, gives the reviewer partial credit, cites what actually strengthens the paper, and hands the decision back to the editor — who is the correct arbiter of citation demands from an anonymous reviewer.

Pitfalls

- **Reviewer 2 demanding extra experiments** unrelated to the paper's scope — decline politely with reasoning; editors will support you.
- **Reviewing under time pressure** — declines are better than perfunctory reviews.
- **Ad hominem comments** in reviews — never.
- **Failure to disclose COI** as a reviewer — as much a problem as failing to disclose as an author.

Self-check

1. A reviewer requests you cite three of their own recent papers. What is the professional response?
2. You are asked to review a paper on a technique you disagree with philosophically. Should you accept?
3. A response letter that starts with 'We thank the reviewer for their insightful comments' is doing what wrong?

Answers

1. Evaluate each on merit. Cite the ones that actually strengthen the paper; explain politely why the others do not fit; let the editor arbitrate. Do not stonewall; do not comply reflexively. See COPE guidance.

2. Only if you can review the methods and data on their own merits. If your disagreement is with the question, decline and say so. If it is with the design or analysis, that is fair game — comment on the design specifically, not on the choice of question.

3. Grovelling wastes both parties' time and dilutes the substantive response. Thank the reviewer once (or not at all) and get to the point-by-point response. Every 'insightful,' 'thoughtful,' 'excellent' costs you credibility.

Anchor references [48,49]

Module 16 · Presentation and risk communication

Learning objectives. After this module you can present a paper at a podium in 10 minutes with the effect size front and center; communicate risk to a patient in a form they can act on; use natural frequencies over percentages when the audience is not statistically trained.

Concept — brief. Two audiences, one principle: **estimate first, threshold second, uncertainty visible.** The evidence base for effective risk communication is now mature and largely converged [50]:

- Use **natural frequencies** ("6 in 100 patients"), not percentages, for lay audiences.
- Show **absolute** risks, not relative alone.
- Use **icon arrays** (10x10 grids of stick figures shaded to show risk) for probability communication.
- Use **decision aids** for preference-sensitive decisions [51].
- Frame both benefit and harm; do not frame one and omit the other.

For patients considering elective arthroplasty — a preference-sensitive decision by definition — shared decision-making improves decision quality and does not measurably worsen surgical outcomes [52]. It also, in a growing evidence base, does not reduce operative volume as much as some fear.

Checklist — podium presentation

- ☐ Title slide, one-sentence question, one-sentence design.
- ☐ Methods in 2 slides maximum.
- ☐ Primary result on one slide — effect size + CI + one interpretation sentence.
- ☐ Subgroups and secondary outcomes in tables or forest plots, not walls of text.
- ☐ Limitations named specifically.
- ☐ Conclusion matches the primary result.
- ☐ 10 minutes = 10 slides max, aim for one point per slide.

Checklist — patient risk communication

- ☐ Natural frequencies with a reference class ("of 100 patients like you…").
- ☐ Absolute risk shown alongside relative.
- ☐ Icon array for probability visualization.
- ☐ Benefit and harm both shown.
- ☐ Uncertainty acknowledged explicitly.
- ☐ Decision aid used for high-stakes elective decisions.

Worked example — arthroplasty.

Explaining PJI risk after primary TKA to a patient with diabetes:

- Bad: “Your risk of infection is about doubled.”
- Better: “In 100 patients without diabetes, about 1 will develop a deep infection in the first year. In 100 patients with well-controlled diabetes like yours, about 2 will. That means 98 patients like you will not have an infection, and 2 will. Here is what we do to lower that number.” Followed by an icon array.

Pitfalls

- **Relative risk alone** for a rare event (“doubles the risk!”) — sounds terrifying, hides absolute magnitude.
- **Framing one side only** — “90% will have no infection” and “10% will have an infection” are the same fact and read differently.
- **Presenter using p-values on slides** without effect sizes — the audience cannot judge the claim.
- **Slides read verbatim** — you become a dispensable narrator.
- **Ignoring the decision-aid literature** and calling that “shared decision-making.”

Self-check

1. You must explain a 2% surgical infection risk to a patient. Which framing is worst?
2. You have 8 minutes at a podium and 32 slides. What do you cut?
3. Why does a decision aid outperform an unstructured conversation for preference-sensitive decisions?

Answers

1. ‘The risk is doubled compared to the general population.’ Relative risk without absolute anchoring is the standard risk-communication anti-pattern. Better: ‘2 in 100 patients like you develop an infection in the first year; 98 do not.’ See Gigerenzer et al. [50].

2. Everything except: one-sentence question, one-slide design, one-slide primary result with effect size + CI, one-slide interpretation, one-slide limitations. If you have 8 minutes, you have 8 slides. The rest is a paper, not a talk.

3. Because it standardizes the presentation of options, probabilities, and outcomes, and reduces the noise from clinician-specific framing. Evidence shows improved decision quality and knowledge without measurable harm to surgical volume. See Jayakumar et al. [51].

Anchor references [50,51,52]

Module 17 · Research ethics and integrity

Learning objectives. After this module you can apply ICMJE authorship criteria; disclose conflicts of interest in a way that protects your work; navigate the emerging norms for LLM/AI use in manuscript preparation.

Concept — brief. Integrity is not a section of the paper — it is the paper. Three durable domains and one emerging one:

- **Authorship:** ICMJE criteria (substantial contribution + drafting/revising + final approval + accountability). Honorary and ghost authorship remain common and remain wrong [53].
- **Conflict of interest:** disclose financial and non-financial COIs. Journal disclosures are minimum; institutional disclosures may be broader [54,55].
- **Research misconduct:** fabrication, falsification, plagiarism. Retractions are rising; the reputational cost is permanent.
- **AI and LLM use:** as of 2025-2026, most major journals require disclosure of any LLM use in writing, prohibit LLM listing as an author, and hold authors accountable for all content including hallucinated citations (see the GTPS reference case in Module 12). Journal policies are converging but not identical; check per submission.

Checklist — ethics and integrity

- ☐ Authorship list reviewed against ICMJE criteria; every author meets all four.
- ☐ Contribution statement (CRediT) written before submission.
- ☐ All financial COIs disclosed to the journal.
- ☐ Non-financial COIs (institutional, ideological) disclosed where journal requires or where they could reasonably influence.
- ☐ IRB or ethics approval documented; consent process described.
- ☐ Registered protocol cited (Module 12).
- ☐ LLM use disclosed per journal policy; every LLM-suggested citation verified against the primary source (see reference-integrity check).
- ☐ Data-sharing plan honored.

Worked example — arthroplasty.

A device-industry-funded RCT of a new stem. Correct handling:

- Full funding disclosure in acknowledgments and methods.
- Statistical analysis performed by an independent statistician not employed by the sponsor, named in the byline or acknowledged.
- Prespecified analysis plan registered; deviations disclosed.
- Authors — including surgeon investigators — retain final control over publication independent of sponsor approval.

- COI disclosures cover consulting, royalties, stock, and travel.
- If any author used an LLM to draft methods text, disclose per journal policy and verify all citations.

Pitfalls

- **"Guest" senior authorship** for a division chief who did not contribute — ICMJE violation.
- **Ghost writing** by a medical-communications agency — undisclosed authorship.
- **Sponsor with veto over publication** — kills the paper's credibility.
- **Undisclosed royalty payments** on a related implant — retraction-grade.
- **Unverified LLM-generated references** — high risk of fabrication; every citation must be checked against the primary source before submission.

Self-check

1. Your industry-funded trial shows a positive effect. What is your obligation in the manuscript?
2. You realize partway through a project that a co-author has a financial COI they did not disclose. What do you do?
3. Your co-author asks you to add an author who did not contribute. What is the correct response?

Answers

1. Disclose the funding source and the sponsor's role in study design, data access, analysis, and manuscript preparation. If the sponsor had veto power over publication, disclose it. Absent disclosure, readers assume the worst. See ICMJE and Fontanarosa & Bauchner [55].

2. Ask them to disclose it to the group and to the journal at submission. If they refuse, that is grounds to remove them from authorship or to withdraw the submission. Undisclosed COI is a retraction-grade defect.

3. Decline and cite ICMJE authorship criteria. Guest authorship (adding a non-contributor to boost prestige or CV) and gift authorship (adding a friend/mentor as a courtesy) are both research-integrity violations. Offer the acknowledgments section if appropriate.

Anchor references [53,54,55]

Module 18 · Calibration, decision journals, and deliberate practice

Learning objectives. After this module you can keep a decision journal that improves your predictions over time; distinguish discrimination from calibration in a prediction model; recognize the cognitive biases that show up in your own decision-making.

Concept — brief. The final layer of the framework — Decision Journal — is a personal discipline, not a study design. It rests on three ideas:

- **Prediction is a testable claim.** If you predict a 20% chance of PJI in a case, you can score yourself against outcomes over time (calibration).
- **Cognitive bias is not a rhetorical flourish.** Anchoring, availability, and confirmation bias are documented in surgical decision-making, and debiasing interventions can reduce them, though not eliminate them [56,57].
- **Calibration is measurable.** In prediction models, calibration ≠ discrimination [45,58]. In your own predictions, calibration = the correlation between your stated probability and observed outcome frequency.

Checklist — decision journal (personal)

- ☐ For each consequential decision, record: date, decision, expected outcome, confidence (probability), reasoning.
- ☐ Set a follow-up date.
- ☐ On follow-up, record actual outcome without editing the original entry.
- ☐ Score calibration quarterly: bin predictions (0-20%, 20-40%, ...); plot predicted vs. observed.
- ☐ Debrief systematic errors: over- or under-confidence, anchoring on the last similar case, availability from recent complications.
- ☐ Update mental models; do not update your journal entries.

Checklist — model calibration (research)

- ☐ Report calibration alongside discrimination (Module 14).
- ☐ Use a calibration curve with confidence band; not just Hosmer-Lemeshow.
- ☐ Report integrated calibration index (ICI) or E-max for a summary measure [45].
- ☐ Recalibrate before external deployment.

Worked example 1 — personal calibration on 90-day readmission.

A surgeon logs predicted 90-day readmission probability for each revision THA. After 100 cases:

- Discrimination (AUC of predicted vs. actual): 0.72 — respectable.
- Calibration: for cases predicted 10-15%, observed rate is 22%. The surgeon systematically underestimates risk in this bin.

- Debrief: probable anchoring on the last two cases (both uneventful). Adjust priors upward for chronic-opioid revisions. Reassess next quarter.

The journal is not for beating yourself up. It is for closing the gap between what you think you know and what you actually know.

Worked example 2 — external validation of a published PJI risk model.

You want to deploy a published preoperative PJI risk model in your practice. In 500 primary TKAs at your center over the last 3 years:

- Discrimination: AUC 0.71 (development cohort reported 0.78) — acceptable degradation.
- Calibration intercept: +0.4 (model systematically underestimates PJI risk in your population by a factor of ±e^0.4 = 1.5x on the odds scale).
- Calibration slope: 0.82 (model over-extremes — high-risk predictions are less high than observed, low-risk less low).
- Integrated calibration index (ICI): 0.03 (moderate miscalibration).

Decision: recalibrate the intercept before local deployment (a one-parameter update on your data). Do not refit the whole model — the discrimination is preserved and refitting on 500 events risks overfit. Add the recalibrated model to the preoperative checklist with a calibration plot posted in the operative planning room so users can see when the risk band applies to a patient like theirs.

The point: a model with AUC 0.71 and ICI 0.03 is deployable after a one-line recalibration. The same model, deployed as-is, would systematically underestimate PJI risk in your patients and quietly under-inform consent.

Worked example 3 — catching an availability-driven prediction error.

Over Q1, you predicted a 15% dislocation rate for a series of complex revision THAs and observed 8%. Over Q2, you predicted 25% and observed 12%. In both quarters, your predictions were systematically too high.

Debrief:

- Q1: A dislocation cluster in December (3 in one week) inflated your priors going into Q1.
- Q2: A high-profile dislocation lawsuit at another center inflated your priors going into Q2.

Mechanism: availability bias driven by recent memorable events, not by base-rate change in your practice.

Corrective actions: (i) at prediction time, force yourself to write the base rate from your own last 200 cases as the anchor before recording a case-specific probability; (ii) require a

written reason for any predicted probability that departs from the base rate by more than 5 percentage points. Reassess calibration next quarter with the debiasing intervention in place; do not retrofit the earlier journal entries.

Pitfalls

- **Editing the journal post-outcome** — destroys the calibration signal.
- **Recording only outcomes you got right** — selection bias in your own record.
- **Confusing discrimination with calibration** — a model can rank patients well and still misestimate their absolute risks.
- **Debiasing training in isolation** — modest effects; systems and structured decision aids do more work than personal effort alone.

Self-check

1. Your decision journal shows you are systematically overconfident in the 20-40% probability band. What is the corrective action?
2. What is the difference between calibration and discrimination in a prediction model, in one sentence each?
3. You externally validate a published model and find good discrimination but poor calibration. What is the minimum acceptable action before local deployment?

Answers

1. Not 'try harder.' Force a structural intervention: at prediction time, write the base rate from a comparable prior series before recording the case-specific probability. Reassess calibration next quarter. See Ludolph & Schulz [57].

2. Discrimination = does the model rank higher-risk patients above lower-risk ones (AUC). Calibration = do the model's predicted probabilities match observed event rates (calibration plot, ICI). A model can rank perfectly and still be badly miscalibrated. See Van Calster et al. [45].

3. Recalibrate — at least intercept recalibration, potentially slope adjustment (logistic recalibration). Do not deploy as-is; do not refit the whole model on a small local cohort. Report the recalibration transparently and post the recalibrated calibration plot alongside the model.

Anchor references [45,56,57,58]

Full reference list — JOA style

1. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak*. 2007;7:16. doi:10.1186/1472-6947-7-16. PMID: 17573961.

2. Willis LD. Formulating the Research Question and Framing the Hypothesis. *Respir Care*. 2023;68(8):1180-1185. doi:10.4187/respcare.10975. PMID: 37041024.

3. Covvey JR, McClendon C, Gionfriddo MR. Back to the basics: Guidance for formulating good research questions. *Res Social Adm Pharm*. 2024;20(1):66-69. doi:10.1016/j.sapharm.2023.09.009. PMID: 37838572.

4. Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. *BMJ*. 2022;378:e071108. doi:10.1136/bmj-2022-071108. PMID: 36041749.

5. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70-75. doi:10.1016/j.jclinepi.2016.04.014. PMID: 27237061.

6. Zuidgeest MGP, Goetz I, Groenwold RHH, Irving E, van Thiel GJMW, Grobbee DE. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *J Clin Epidemiol*. 2017;88:7-13. doi:10.1016/j.jclinepi.2016.12.023. PMID: 28549929.

7. Järvinen TL, Sihvonen R, Bhandari M, Sprague S, Malmivaara A, Paavola M, Schünemann HJ, Guyatt GH. Blinded interpretation of study results can feasibly and effectively diminish interpretation bias. *J Clin Epidemiol*. 2014;67(7):769-772. doi:10.1016/j.jclinepi.2013.11.011. PMID: 24560088.

8. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. *J Am Med Inform Assoc*. 2023;30(5):971-977. doi:10.1093/jamia/ocad013. PMID: 36752649.

9. Inoue K, Sakamaki K, Komukai S, Ito Y, Goto A, Shinozaki T. Methodological Tutorial Series for Epidemiological Studies: Confounder Selection and Sensitivity Analyses to Unmeasured Confounding From Epidemiological and Statistical Perspectives. *J Epidemiol*. 2025;35(1):3-10. doi:10.2188/jea.JE20240082. PMID: 38972732.

10. Gagnier JJ. Patient reported outcomes in orthopaedics. *J Orthop Res*. 2017;35(10):2098-2108. doi:10.1002/jor.23604. PMID: 28513993.

11. Rolfson O, Eresian Chenok K, Bohm E, Lübbeke A, Denissen G, Dunn J, Lyman S, Franklin P, Dunbar M, Overgaard S, Garellick G, Dawson J; Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries. Patient-reported outcome measures in arthroplasty registries. *Acta Orthop*. 2016;87(Suppl 1):3-8. doi:10.1080/17453674.2016.1181815. PMID: 27168175.

12. Reynaud V, Verdilos A, Pereira B, Boisgard S, Costes F, Coudeyre E. Core Outcome Measurement Instruments for Clinical Trials of Total Knee Arthroplasty: A

Systematic Review. *J Clin Med*. 2020;9(8):2439. doi:10.3390/jcm9082439. PMID: 32751523.

13. Saville BR, Detry MA, Viele K. Conditional Power: How Likely Is Trial Success? *JAMA*. 2023;329(6):508-509. doi:10.1001/jama.2022.25080. PMID: 36689237.

14. Kunselman AR. A brief overview of pilot studies and their sample size justification. *Fertil Steril*. 2024;121(6):899-901. doi:10.1016/j.fertnstert.2024.01.040. PMID: 38331310.

15. Wang X, Ji X. Sample Size Estimation in Clinical Research: From Randomized Controlled Trials to Observational Studies. *Chest*. 2020;158(1S):S12-S20. doi:10.1016/j.chest.2020.03.010. PMID: 32658647.

16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) — a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010. PMID: 18929686.

17. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN; REDCap Consortium. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208. PMID: 31078660.

18. Pellen C, Munung NS, Armond AC, Bruckner T, Cristea IA, Danchev V, Dechartres A, Erlichman J, Goldacre B, Hemkens LG, Hopewell S, Ioannidis JPA, Kaltenegger L, Kotronoulas G, Lambert P, Landrum M, Lawrence-Jones A, Lee K, Marušić A, Mayo-Wilson E, Moher D, Naudet F, Norman-Bruce H, Ohmann C, Ozalp E, Ross JS, Schriger DL, Serghiou S, Turner E, Vandvik PO, Wager E, Boutron I. Data management and sharing. *J Clin Epidemiol*. 2025;180:111680. doi:10.1016/j.jclinepi.2025.111680. PMID: 39842522.

19. Kallogjeri D, Piccirillo JF. A Simple Guide to Effect Size Measures. *JAMA Otolaryngol Head Neck Surg*. 2023;149(5):447-451. doi:10.1001/jamaoto.2023.0159. PMID: 36951858.

20. Madjarova SJ, Williams RJ 3rd, Nwachukwu BU, Martin RK, Karlsson J, Ollivier M, Pareek A, Nakamura N. Picking apart p values: common problems and points of confusion. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(10):3245-3248. doi:10.1007/s00167-022-07083-3. PMID: 35920843.

21. Austin PC, White IR, Lee DS, van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can J Cardiol*. 2021;37(9):1322-1331. doi:10.1016/j.cjca.2020.11.010. PMID: 33276049.

22. Polmear M, Vasilopoulos T, O'Hara N, Wooster BM, Zeini I. Death of the P Value? Bayesian Statistics for Orthopaedic Surgeons. *J Am Acad Orthop Surg*. 2025;33(6):285-300. doi:10.5435/JAAOS-D-24-00813. PMID: 39661776.

23. See reference 21 above — duplicate source; retained here to preserve original in-text citation numbering. *(Missing data / multiple imputation.)*

24. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA*. 2022;327(11):1083-1084. doi:10.1001/jama.2022.1816. PMID: 35226050.

25. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, Harrison WJ, Keeble C, Ranker LR, Textor J, Tomova GD, Gilthorpe MS, Ellison GTH. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620-632. doi:10.1093/ije/dyaa213. PMID: 33330936.

26. Jafarzadeh SR, Neogi T. Causal inference from observational data and target trial emulation. *Osteoarthritis Cartilage*. 2022;30(11):1415-1417. doi:10.1016/j.joca.2022.08.010. PMID: 36063988.

27. See reference 5 above — duplicate source; retained here to preserve original in-text citation numbering. *(Hernán target trial specification.)*

28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Med*. 2021;18(3):e1003583. doi:10.1371/journal.pmed.1003583. PMID: 33780438.

29. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, McKenzie JE. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021;372:n160. doi:10.1136/bmj.n160. PMID: 33781993.

30. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JPA, Straus S, Thorlund K, Jansen JP, Mulrow C, Catalá-López F, Gøtzsche PC, Dickersin K, Boutron I, Altman DG, Moher D. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med*. 2015;162(11):777-784. doi:10.7326/M14-2385. PMID: 26030634.

31. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev*. 2017;6(1):245. doi:10.1186/s13643-017-0644-y. PMID: 29208034.

32. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng HY, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898. PMID: 31462531.

33. Guyatt G, Agoritsas T, Brignardello-Petersen R, Iorio A, Sultan S, Yepes-Nuñez JJ, Murad MH, Kahale L, Alper B, Amer YS, Chen Y, Coello PA, Craigie S, Cuello-Garcia C, Djulbegovic B, Foroutan F, Granholm A, Guyatt J, Hazlewood G, Hoogeveen J, Klugar M, Kredo T, Meerpohl JJ, Mustafa RA, Neumann I, Nieuwlaat R, Oczkowski S, Piggott T, Qaseem A, Rada G, Rojas-Reyes MX, Santesso N, Schünemann HJ, Siemieniuk RAC, Skoetz N, Sun X, Thomas J, Vandvik PO, Vist G, Xie F, Alonso-Coello P, Akl EA. Core GRADE 1: overview of the Core GRADE approach. *BMJ*. 2025;389:e081903. doi:10.1136/bmj-2024-081903. PMID: 40262844.

34. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Del Giovane C, Egger M, Salanti G. CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLoS Med*. 2020;17(4):e1003082. doi:10.1371/journal.pmed.1003082. PMID: 32243458.

35. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RCW, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JPA, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegfried NL, Staniszewska S, Taylor RS, Thabane L, Torgerson D, Vohra S, White IR, Boutron I. CONSORT 2025 Statement: Updated Guideline for Reporting Randomized Trials. *JAMA*. 2025;333(22):1998-2005. doi:10.1001/jama.2025.4347. PMID: 40228499.

36. Chan AW, Boutron I, Hopewell S, Moher D, Schulz KF, Collins GS, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RCW, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JPA, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegfried NL, Staniszewska S, Taylor RS, Thabane L, Torgerson D, Vohra S, White IR, Hróbjartsson A. SPIRIT 2025 Statement: Updated Guideline for Protocols of Randomized Trials. *JAMA*. 2025;334(5):435-443. doi:10.1001/jama.2025.4486. PMID: 40294593.

37. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use

regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378. PMID: 38626948.

38. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x. PMID: 32908283.

39. Turkiewicz A, Luta G, Hughes HV, Ranstam J. Statistical mistakes and how to avoid them — lessons learned from the reproducibility crisis. *Osteoarthritis Cartilage*. 2018;26(11):1409-1411. doi:10.1016/j.joca.2018.07.017. PMID: 30096356.

40. Mathur MB, Fox MP. Toward Open and Reproducible Epidemiology. *Am J Epidemiol*. 2023;192(4):658-664. doi:10.1093/aje/kwad007. PMID: 36627249.

41. Rockhold F, Bromley C, Wagner EK, Buyse M. Open science: The open clinical trials data journey. *Clin Trials*. 2019;16(5):539-546. doi:10.1177/1740774519865512. PMID: 31347390.

42. Knottnerus JA, Tugwell P. How to write a research paper. *J Clin Epidemiol*. 2013;66(4):353-354. doi:10.1016/j.jclinepi.2013.01.007. PMID: 23452724.

43. Willis LD. How to Write the Methods Section of a Research Manuscript. *Respir Care*. 2023;68(12):1763-1770. doi:10.4187/respcare.11437. PMID: 37643870.

44. Note on AI-generated text detection: journal-policy-level guidance is evolving rapidly; see reference 55 for representative COI/authorship framework. Verify current policy at submission time; no single anchor reference is stable enough to cite for LLM detection specifically.

45. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005. PMID: 26772608.

46. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA*. 2015;313(4):409-410. doi:10.1001/jama.2015.37. PMID: 25626037.

47. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol*. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038. PMID: 30241973.

48. Welsby PD. Peer reviews. A peer reviewer's view. *Postgrad Med J*. 2020;96(1140):725-727. doi:10.1136/postgradmedj-2020-138793. PMID: 32943475.

49. Zimba O, Gasparyan AY. Peer review guidance: a primer for researchers. *Reumatologia*. 2021;59(1):3-8. doi:10.5114/reum.2021.102709. PMID: 33707789.

50. Zipkin DA, Umscheid CA, Keating NL, Allen E, Aung K, Beyth R, Kaatz S, Mann DM, Sussman JB, Korenstein D, Schardt C, Nagi A, Sloane R, Feldstein DA. Evidence-based risk communication: a systematic review. *Ann Intern Med*. 2014;161(4):270-280. doi:10.7326/M14-0295. PMID: 25133362.

51. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, Aksan N, Rathouz PJ, Bozic KJ. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults With Knee Osteoarthritis: A Randomized Clinical Trial. *JAMA Netw Open*. 2021;4(2):e2037107. doi:10.1001/jamanetworkopen.2020.37107. PMID: 33599773.

52. Wilson CD, Probe RA. Shared Decision-making in Orthopaedic Surgery. *J Am Acad Orthop Surg*. 2020;28(23):e1032-e1041. doi:10.5435/JAAOS-D-20-00556. PMID: 32925380.

53. Wislar JS, Flanagin A, Fontanarosa PB, Deangelis CD. Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey. *BMJ*. 2011;343:d6128. doi:10.1136/bmj.d6128. PMID: 22028479.

54. Grundy Q, Dunn AG, Bero L. Improving researchers' conflict of interest declarations. *BMJ*. 2020;368:m422. doi:10.1136/bmj.m422. PMID: 32161006.

55. Fontanarosa P, Bauchner H. Conflict of Interest and Medical Journals. *JAMA*. 2017;317(17):1768-1771. doi:10.1001/jama.2017.4563. PMID: 28464125.

56. Aylmore H, Agarwal S, Marcus HJ, Kolias AG, Hutchinson PJ, Gupta S. Cognitive Biases and Heuristics in Surgical Settings: A Systematic Review. *Ann Surg*. 2025;282(6):946-953. doi:10.1097/SLA.0000000000006736. PMID: 40255176.

57. Ludolph R, Schulz PJ. Debiasing Health-Related Judgments and Decision Making: A Systematic Review. *Med Decis Making*. 2018;38(1):3-13. doi:10.1177/0272989X17716672. PMID: 28649904.

58. See reference 45 above — duplicate source; retained here to preserve original in-text citation numbering. *(Calibration hierarchy for risk models.)*

Foundational works not cited above but worth owning

The course above is a working scaffold, not a canon. If you are building a research career, own these in full:

- **Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors.** *Cochrane Handbook for Systematic Reviews of Interventions*. Current version at https://training.cochrane.org/handbook.

- **Hernán MA, Robins JM.** *Causal Inference: What If*. Chapman & Hall/CRC, 2020. Free PDF at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
- **Rothman KJ, Greenland S, Lash TL, editors.** *Modern Epidemiology*. 4th ed. Wolters Kluwer, 2021.
- **Steyerberg EW.** *Clinical Prediction Models*. 2nd ed. Springer, 2019.
- **ICMJE Recommendations** — http://www.icmje.org/. Updated periodically.
- **EQUATOR Network** — https://www.equator-network.org/. The authoritative index of current reporting guidelines.

A closing note on this course

This is a v1 scaffold. It is dense enough to teach from, checklist-first enough to work from, and referenced enough to defend. It is not exhaustive, and it is not a substitute for the foundational texts above. Modules will be revised, examples deepened, and references updated as the OrthoScienus Teaching layer expands.

If a claim in this document contradicts your primary source, trust the primary source and tell me where the drift occurred.

— Jack Walsh, DO