23 The research workflow

R. Heller · Date: 2026-05-08 · Last changed: 2026-10-30

APPENDIX · RESEARCH WORKFLOW

Nine stages from a research question to knowledge that survives replication.

Statistics is a thread running through a much longer process. The diagram below names the stages of that process and links each to the parts of the curriculum that cover it in detail. No stage is optional; errors made at one are paid for at the next.

```
flowchart TD
  Q[1. Question] ⟶ M[2. Measurements]
  M ⟶ D[3. Design]
  D ⟶ A[4. Acquisition]
  A ⟶ Desc[5. Description]
  Desc ⟶ An[6. Analysis]
  An ⟶ I[7. Interpretation / Prediction]
  I ⟶ V[8. Validation]
  V ⟶ K[9. Knowledge / Decisions]
```

23.1 1. Question

A well-formed research question has a population, an exposure or intervention, a comparator, an outcome, and a time frame. The PICO framework from evidence-based medicine is a useful checklist. A vague question leads to a vague design and a vague answer; a well-formed question survives peer review before a single datum is collected. *Covered in:* Scientific process and research workflow, Systematic reviews and prisma.

23.2 2. Measurements

Every question depends on the measurement scale you choose to answer it. Accuracy, precision, reliability, and the difference between a measurement and a proxy are the vocabulary. *Covered in:* Data types tidy data accuracy and precision, Kappa icc bland altman.

23.3 3. Design

Design is where you choose what the data can and cannot tell you. A well-designed observational study answers some questions better than a badly designed RCT; a good RCT answers them all. *Covered in:* Observational designs and strobe.

23.4 4. Acquisition

Getting the data into a computer faithfully is itself a statistical problem: missingness, measurement error, and batch effects start here. *Covered in:* Import joins and missingness with dplyr, Mcar mar mnar.

23.5 5. Description

Before any inference, look at the data. A well-designed plot is often the whole answer. *Covered in:* Ggplot2 grammar and multi panel layouts, Descriptive statistics and table 1.

23.6 6. Analysis

The formal model that takes the data and returns an estimate, an interval, and a decision. *Covered in:* Courses 1–4, throughout.

23.7 7. Interpretation / prediction

What does the estimate mean in the scale of the science? Effect size, clinical significance, and the distinction between an average and an individual prediction live here. *Covered in:* Calibration discrimination roc auc brier score, Time dependent brier ipa external validation.

23.8 8. Validation

Does the finding hold in data not used to find it? External validation, nested CV, and replication pool here. *Covered in:* Cross validation nested cv bootstrap 632, Time dependent brier ipa external validation.

23.9 9. Knowledge / decisions

The act of writing it down so that others can trust, cite, and build on the finding. *Covered in:* Writing a report, Explanation vs prediction reporting.