WEAVE SPRINT · PAPER I

# Research Methods, Biostatistics & Epidemiology

*The whole research and statistics toolkit of the first paper, in one document. Three bands: research methodology, biostatistics and epidemiology, running from the evidence pyramid and the study designs through the significance tests, the diagnostic 2x2 and statistical inference to the measures of disease frequency and the National Mental Health Survey. Built so you can pick the right design for a question and the right test for the data.*

CONTENTS


- Research methodology
- Biostatistics
- Epidemiology
- Discriminate
- Apply and consolidate



Dr. Wilfred D'souza · Dr. Niharika Reddy

WEAVE · CENTRE FOR INTEGRATIVE PSYCHIATRY

Weave
Centre for Integrative Psychiatry

WEAVE SPRINT
PSYCHIATRY OUTREACH PROGRAMME

# Psychiatry that listens before it prescribes.

*Therapy-informed psychiatry for adults, children, couples and families. Online across India, in person in Mumbai.*





## Assessments

ADHD, autism and developmental. Anxiety, mood, attention and personality patterns.



## Consultations and sessions

Psychiatric consultations, therapy sessions, medication review and couples sessions.



## Family Library

Free plain-language guides for families, in five languages.

`weave.clinic/family`

REFER, OR SEND SOMEONE OUR WAY

## Your patient contacts Weave directly and decides whether to book.

You remain the referring clinician. Any coordination between clinicians happens with the patient's permission.

CONTACT

`weave.clinic/contact`
`WhatsApp +91 7625004229`

Dr. Niharika Reddy and Dr. Wilfred D'souza, Consultant Psychiatrists. Runwal Forests Clinic, Kanjurmarg West, Mumbai, and online.

WEAVE.CLINIC

Routine referrals only. This is not an emergency route.

# Contents

Concept / rule  Key number  Trap


**Research methodology** 4

1 Orientation: evidence-based medicine and the hierarchy of evidence 4

2 Observational study designs 10

3 The randomised controlled trial and drug-trial phases 18

4 Validity, bias and confounding 23

5 Sampling methods 30

6 Qualitative research methods 36

7 Systematic review and meta-analysis 41

8 Reporting guidelines and research ethics 47

**Biostatistics** 52

9 Data types and scales of measurement 52

10 Describing data: central tendency, dispersion and the normal curve 57

11 Tests of significance: parametric and non-parametric 62

12 Correlation, regression and factor analysis 67

13 Diagnostic test statistics 73

14 Statistical inference: p-values, confidence intervals, errors and power 79

15 Survival analysis, NNT and ROC 84

**Epidemiology** 89

16 Measures of disease frequency and association 89

17 The National Mental Health Survey and the Indian picture 95

18 Principles of screening 99

**Discriminate** 104

19 Discriminate: pick the design, pick the test 104

**Apply and consolidate** 111

20 Apply: four worked appraisal vignettes 111

21 Consolidate: mnemonic, retrieval and synthesis 116

**Previously asked** 121

**Quick review** 123

**References** 125

**Index** 127

PAPER I

# Research Methods, Biostatistics & Epidemiology

Five sections · 21 topics · one complete notes document

This material gathers the whole research and statistics toolkit of the first paper into one place, because the examiner tests it as one skill: the ability to read a study and reason about numbers. The first band is **research methodology**, the architecture of clinical evidence: the hierarchy that ranks a case report below a cohort below a randomised trial below a systematic review, the three observational designs and their measures of association, the randomised controlled trial with its machinery of randomisation, allocation concealment, blinding and intention-to-treat analysis, the taxonomy of validity and bias and confounding, the sampling methods, the qualitative approaches, the systematic review and meta-analysis with its forest plot, and the reporting guidelines and research ethics that govern the whole enterprise. The second band is **biostatistics**, the grammar of the numbers: the scales of measurement that decide which test is legal, the description of data and the normal curve, the parametric and non-parametric significance tests, correlation and regression, the diagnostic test statistics built on the 2x2 table, statistical inference with its p-values and confidence intervals and errors, and survival and ROC analysis. The third band is **epidemiology**: incidence and prevalence, relative risk and the odds ratio, the National Mental Health Survey numbers that anchor the Indian picture, and the principles of screening. The examinable skill is judgement on both fronts: choosing the right design for a question, choosing the right test for the data, and interpreting the result correctly.

**HOW TO USE THESE NOTES**

Read the three bands as one argument. In **methodology**, every named design ends with a **Good to remember** box giving what it is, when to use it, and the one number or feature to know, so the reference facts sit apart from the reasoning. In **biostatistics**, the same box holds each test's one-line rule: what data it takes and what it answers. The **Discriminate** section is the payoff, two selector tables, one to choose a study design from the question and one to choose a statistical test from the data, and the **Apply** section works four short appraisal cases end to end. Figures declare one axis and code it in colour and shape: **petrol** marks structure and the neutral case, **coral** marks the trap or the point of discrimination, and **marigold** highlights the number to hold. Two boundaries are worth stating up front: the twin, family and adoption designs as genetic methods live in the Psychiatric Genetics notes, and the epidemiology of specific disorders lives in the clinical chapters of Paper II, while this material owns the general research toolkit and the national survey numbers; and the reliability and validity of a study here are a different idea from the psychometric reliability and validity of a test in the Psychometry, Assessment and Descriptive Psychopathology notes. This is doctor-facing revision, so the full technical vocabulary is used.

# Research methodology

## 1 · Orientation: evidence-based medicine and the hierarchy of evidence

These notes cover one examinable skill wearing two hats: reading a study and reasoning about its numbers. That is all research methods and statistics really are for the psychiatrist. A paper arrives; you must place its **design**, judge whether its **data** were handled correctly, and decide whether its estimate of a **disease frequency** or a treatment effect can be trusted enough to change what you do on Monday. Methods and statistics are not two subjects but the front and back of the same coin, and the examiner tests them together: a viva that hands you an abstract wants to hear you name the design, state what it can and cannot answer, spot the trap, and read the confidence interval.

**Why this opens the notes.** Before any single test or design can be defended, you need the map that ranks them, because the whole discipline is organised around one idea: some kinds of evidence are more trustworthy than others, and the trustworthiness follows from the design, not from the conclusion. That map is the **study-design hierarchy**, the evidence pyramid of evidence-based medicine. Learn the pyramid first and every later topic, from the randomised controlled trial to the odds ratio to the funnel plot, hangs on a rung you already recognise (Cipriani, Leucht & Geddes 2020).

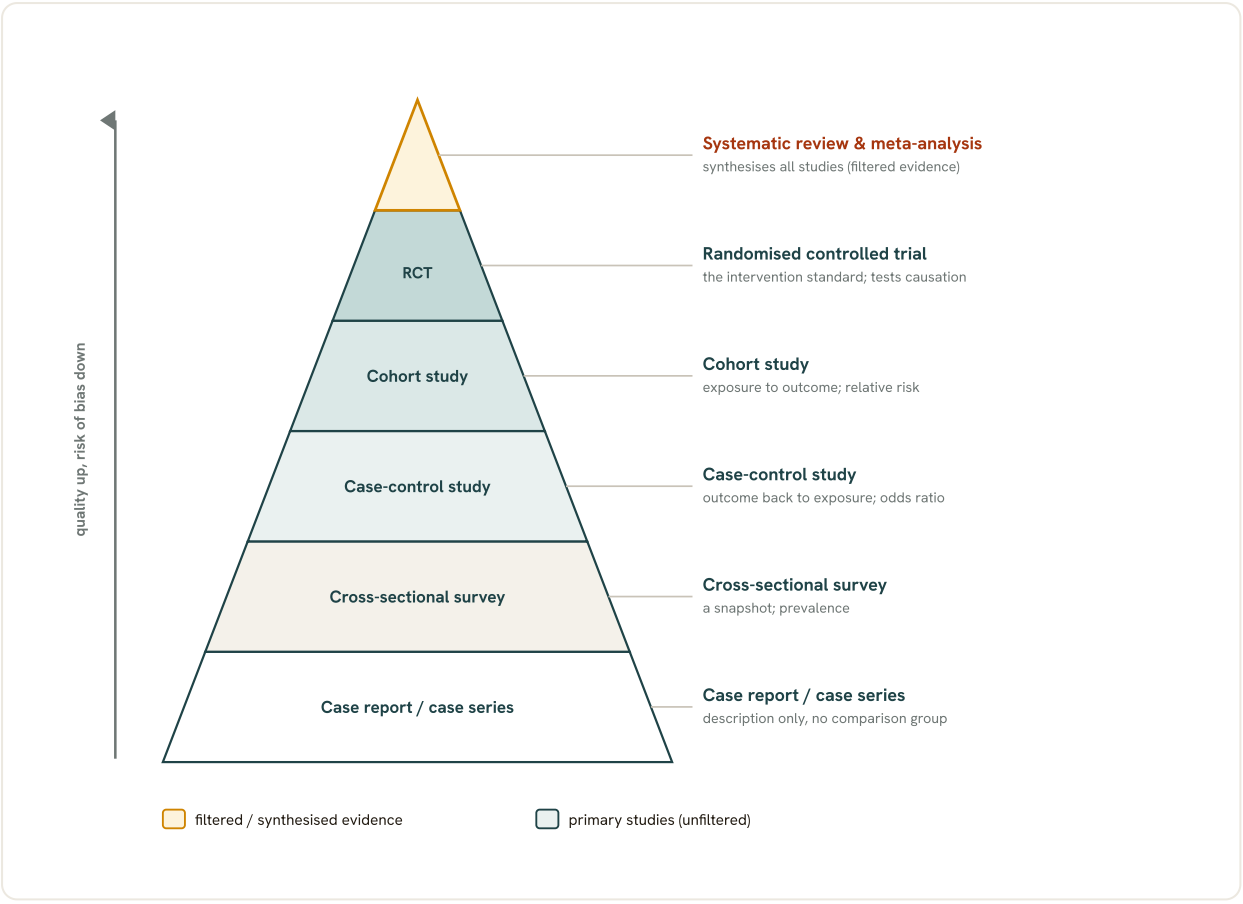



**Fig 7.1** The hierarchy of evidence. Study designs rank by how well they control bias: the case report at the base offers only description, the observational designs add comparison groups, the randomised controlled trial adds the control of randomisation, and the systematic review with meta-analysis at the apex synthesises the whole body of filtered evidence. Reading a study begins with placing it on this pyramid, because the design sets the ceiling on the strength of the conclusion.

## 1.1 Evidence-based medicine: the definition and why it is more than "reading trials"

**Definition.** Evidence-based medicine is the conscientious integration of best research evidence with clinical expertise and patient values (Sackett et al. 1996). Three legs, not one: the *best current research evidence* from methodologically sound studies, the *clinician's own expertise* built from cumulative experience and skill, and the *patient's preferences, concerns and values*. Drop any leg and the stool falls: pure evidence with no clinical judgement is cookbook medicine, pure expertise with no evidence is eminence-based medicine, and neither is defensible. The term **ebm** in an exam answer should always cue all three components, because a common one-mark question is simply "define evidence-based medicine and name its components".

RULE **Three legs or none.** If your answer names research evidence but omits clinical expertise and patient values, it is not a definition of ebm.

## 1.2 The five steps of practising ebm (the 5 A's)

**The cycle.** Practising evidence-based medicine is a repeatable loop. The classic account gives four steps of doing plus a fifth of self-audit, usually taught as the **five steps** or 5 A's. These are the **5 steps** the examiner wants listed in order:

### 1. Ask

Convert the clinical problem into a focused, answerable question (the PICO question, 1.3).

**2. Acquire**
Search efficiently for the best available evidence, choosing the right resource (systematic reviews before single trials).

**3. Appraise**
Critically appraise the evidence for validity (closeness to the truth), size of effect, and applicability.

**4. Apply**
Integrate the appraised evidence with clinical expertise and the patient's values, and act.

**5. Assess**
Evaluate your own performance in steps 1 to 4 and the outcome in this patient, then refine.

Steps 1.2 through the rest of these notes are really an expansion of the middle three A's: how to frame the question, how to rank and read the designs you acquire, and how to appraise the numbers.

**WORKED EXAMPLE**

A patient with treatment-resistant depression asks about adding lithium. **Ask:** in adults with unipolar depression not responding to an antidepressant (P), does adjunctive lithium (I) versus continued antidepressant alone (C) improve response (O)? **Acquire:** a systematic review of augmentation trials sits above any single trial. **Appraise:** check randomisation, blinding, effect size and confidence interval. **Apply:** weigh the benefit against monitoring burden and the patient's willingness to have blood levels checked. **Assess:** review response and tolerability at follow-up. That is one full turn of the five steps.

### 1.3 The focused question: PICO

**Framing.** A vague question ("does lithium help?") cannot be searched or answered; a focused one can. The standard scaffold for a therapy or aetiology research question is PICO: **P**opulation (which patients), **I**ntervention (or exposure), **C**omparison (against what, often placebo or usual care), and **O**utcome (measured how). The word **pico** is the examiner's cue for "how do you build an answerable clinical question", and a well-built PICO also tells you which design and which search terms you need. For a harm or prognosis question the I becomes an exposure and the C its absence, but the four slots hold.

TRAP **The missing comparator.** A "study" with no C slot ("patients on drug X improved") is a case series, not a comparative study, and cannot establish that X did the improving.

### 1.4 The hypothesis: why we test the null

**Setting up the test.** Once the PICO question is framed, the statistics needs it as a testable hypothesis. Two hypotheses are stated as a pair. The null hypothesis (H0) is the hypothesis of *no difference* between the groups, or *no association* between exposure and outcome; the alternative hypothesis (H1) is that a difference or association does exist. The logic of frequentist testing is deliberately conservative: we never "prove" the alternative, we assume the null is true, compute how surprising the observed data would be under it, and reject the null only if that probability (the p-value) is small enough (Bland 2015).

**Why the null, and not the alternative.** The null is a single, precise statement ("the two means are equal") whose sampling behaviour we can calculate exactly; the alternative is a whole family of possibilities ("they differ, by some unknown amount") that we cannot. So we test the tractable one. Rejecting H0 lets us *infer* H1; failing to reject H0 is not proof of H0, only absence of evidence against it. This asymmetry, assume no difference and demand the data disprove it, is the spine of null hypothesis significance testing (Altman 1991).

HOLD THIS

We test the null hypothesis of "no difference" because it is the one precise, calculable statement; a small p-value lets us reject it and infer the alternative, but failing to reject it never proves the null.

### 1.5 The study-design hierarchy: the evidence pyramid

**The map.** Studies are ranked into a hierarchy of evidence, drawn as a pyramid, because design determines how well a study controls bias and confounding, and therefore how close its estimate sits to the truth. From the broad, weak base up to the narrow, strong apex the study-design hierarchy runs: the case report and case series, then the cross-sectional survey, then the case-control study, then the cohort study, then the randomised controlled trial, and at the top the systematic review and meta-analysis. The base is wide because weak evidence is abundant and easy to generate; the apex is narrow because strong evidence is scarce and expensive. Only the designs near the top are called "best evidence" for a question of treatment efficacy, where the randomised controlled trial reigns because randomisation is the only known way to balance both known and unknown confounders (Cipriani, Leucht & Geddes 2020).

**Filtered versus unfiltered.** A second cut runs across the pyramid. The single studies (case report through randomised controlled trial) are *unfiltered* or primary evidence: you must appraise each one yourself. The top band, systematic reviews and meta-analyses, is *filtered* or secondary evidence: someone has already gathered, appraised and (in a meta-analysis) statistically pooled the primary studies for you. This is why an efficient search in step 2 starts at the top and works down only if nothing filtered exists.

RULE **Match the design to the question.** The pyramid ranks designs for *treatment efficacy*; for a rare adverse effect a case-control study, not a randomised controlled trial, may be the best feasible design.

### 1.6 Levels of evidence: Oxford CEBM and GRADE

**From pyramid to grade.** The pyramid is a picture; formal schemes turn it into ranked levels of evidence. The Oxford Centre for Evidence-Based Medicine (Oxford CEBM) scheme assigns levels by design and question type, with systematic reviews of randomised trials at the top level and expert opinion at the bottom. GRADE takes a different route: rather than ranking by design alone, it starts each body of evidence high (for trials) or low (for observational studies) and then rates it up or down across four certainty grades, high, moderate, low and very low, according to risk of bias, inconsistency, indirectness, imprecision and publication bias. The examinable contrast is that Oxford CEBM ranks *studies* by design, whereas GRADE rates the *certainty of the*

*whole body of evidence* for a specific outcome, which is why the same trial can support a high or a low GRADE depending on the other studies and the outcome asked about.

**INDIA**

Indian clinical practice guidelines, including those of the Indian Psychiatric Society, adopt GRADE-style certainty ratings and the same evidence hierarchy, so the language of levels of evidence is not academic decoration: it is how a national guideline tells you the strength behind each recommendation you are expected to follow in a viva on management.

### 1.7 Systematic review and meta-analysis: the apex, and why it is not automatic

**The top band.** A systematic review answers a narrow, pre-specified question by an explicit, reproducible method: define the question, search comprehensively, apply stated inclusion criteria, appraise every included study, and synthesise. A *meta-analysis* is the optional statistical step inside a systematic review that pools the primary studies into one summary effect size with greater power than any single trial (Cipriani, Leucht & Geddes 2020). Not every systematic review contains a meta-analysis, and pooling is only justified when the studies are similar enough to combine.

**The apex can still mislead.** A meta-analysis inherits the flaws of its inputs. Its most serious threat is publication bias: if positive trials are published and negative ones buried, pooling the survivors overstates the effect. This is why the apex is a rung on a ladder, not a guarantee, and appraisal never stops just because a paper says "systematic review" in the title.

**TRAP** **Apex worship.** "It is a meta-analysis, so it is true" ignores publication bias and heterogeneity; a meta-analysis of biased trials produces a precise, confident, wrong answer.

### 1.8 The master mnemonic: design, then data, then disease-frequency

**The spine of everything that follows.** These notes move through three bands in order, and one chained mnemonic holds the whole arc: read the **design** first, then check the **data** handling, then interpret the **disease-frequency** (or effect) it estimates. Every topic is a D in that chain. The pyramid you have just learned is the design band; the statistics topics are the data band; the epidemiology topics are the disease-frequency band.

**MNEMONIC**

#### THREE D's: Design, Data, Disease-frequency

The three bands of these notes, in reading order. **Design**: place the study on the evidence pyramid and know what its rung can answer. **Data**: was the null tested correctly, the right statistical test chosen, the p-value and confidence interval read honestly. **Disease-frequency**: what does it say about incidence, prevalence, risk or effect, and can that be generalised. Name the design, trust the data, then quote the frequency: any exam abstract can be attacked in that order.

### 1.9 What each design answers and where it sits

**The catalogue in one view.** Before the detailed topics, fix each rung with the question it answers and the one measure it yields. This table is the reference you will return to whenever an abstract lands.

| DESIGN (BASE TO APEX) | WHAT IT ANSWERS | SIGNATURE MEASURE | DIRECTION / TIME |
|---|---|---|---|
| **Case report / series** | Does this phenomenon exist at all; hypothesis generation | **Description only, no rate** | None (no comparator) |
| **Cross-sectional** | How common is it now; is there an association at one time point | **Prevalence** | Snapshot (one moment) |
| **Case-control** | What exposures differ between those with and without the disease (good for rare disease) | **Odds ratio** | Backward (outcome to exposure) |
| **Cohort** | Do the exposed develop the outcome more than the unexposed (good for rare exposure) | **Relative risk, incidence** | Forward (exposure to outcome) |
| **Randomised controlled trial** | Does the intervention cause the outcome (efficacy) | **Relative risk, NNT** | Forward, allocation randomised |
| **Systematic review / meta-analysis** | What does all the evidence together say | **Pooled effect size** | Synthesis of the above |

The evidence pyramid as a table: each rung with the one question it answers and the one measure it yields. The odds ratio, relative risk and NNT are unpacked in their own topics; here, hold which design gives which.

**GOOD TO REMEMBER**

- **Case report / series:** a description of one or a few patients with no comparison group; use to flag a novel presentation or generate a hypothesis; gives no rate at all.
- **Cross-sectional study:** measures exposure and outcome at a single time point in a defined population; use to estimate how common something is; the one number is prevalence = cases / total population at that time.
- **Case-control study:** starts from the outcome and looks back at past exposures, comparing cases with controls; best for rare diseases; the one measure is the odds ratio = odds of exposure in cases / odds of exposure in controls.
- **Cohort study:** follows exposed versus unexposed forward in time to see who develops the outcome; best for rare exposures and gives incidence; relative risk = incidence in exposed / incidence in unexposed.
- **Randomised controlled trial:** allocates the intervention by a concealed random method to balance confounders; the top design for efficacy; yields relative risk and number needed to treat, NNT = 1 / absolute risk reduction.
- **Systematic review and meta-analysis:** an explicit, reproducible synthesis of primary studies, optionally pooling them statistically; the apex and filtered evidence; the meta-analysis gives one pooled effect size with tighter confidence intervals than any single study.
- **Oxford CEBM levels:** a scheme that ranks individual studies into levels by design and question type; use to grade a single paper; systematic reviews of trials sit at the top level.
- **GRADE:** rates the certainty of a whole body of evidence per outcome into four grades; use in guideline-making; starts high for trials, low for observational studies, then moves up or down.

**PEARL**

Two axes cross the pyramid and both are examinable. The *vertical* axis is trustworthiness by design (case report at the base, meta-analysis at the apex). The *horizontal* axis is filtered versus unfiltered: everything from case report to randomised controlled trial is unfiltered primary evidence you must appraise yourself, while systematic reviews and meta-analyses are filtered secondary evidence pre-appraised for you. Say both axes and you have shown you understand the pyramid, not just memorised its order.

For genetic study designs such as twin, family and adoption studies, which sit alongside the cohort and case-control designs but answer questions of heritability, see the Psychiatric Genetics notes. For the epidemiology of specific disorders and the national prevalence figures, see the clinical chapters of Paper II. The reliability and validity discussed here belong to *study* quality; the distinct psychometric reliability and validity of a *test* are covered in the Psychometry, Assessment and Descriptive Psychopathology notes.

## 2 · Observational study designs

**This is the map every research question is read against.** In an **observational** study the investigator watches nature take its course and measures exposures and outcomes, but does not allocate the exposure. That single fact separates the whole family from the interventional trial (where the investigator assigns the exposure) and is the first fork in any study-design question. The examiner wants you to name a design from a one-line vignette, state its **direction of enquiry** (does it start from exposure or from outcome?), and give the correct measure of

association it yields. Get those three right and almost every stem falls open (Bland 2015; Altman 1991).

The three analytic **observational** designs (cross-sectional, case-control, cohort) differ chiefly in their **time axis**: the cross-sectional design takes a single snapshot, the case-control design looks backward from outcome to exposure, and the cohort design follows forward from exposure to outcome. The ecological study sits apart because its unit of analysis is the group, not the individual. Hold the time axis and the rest is bookkeeping.

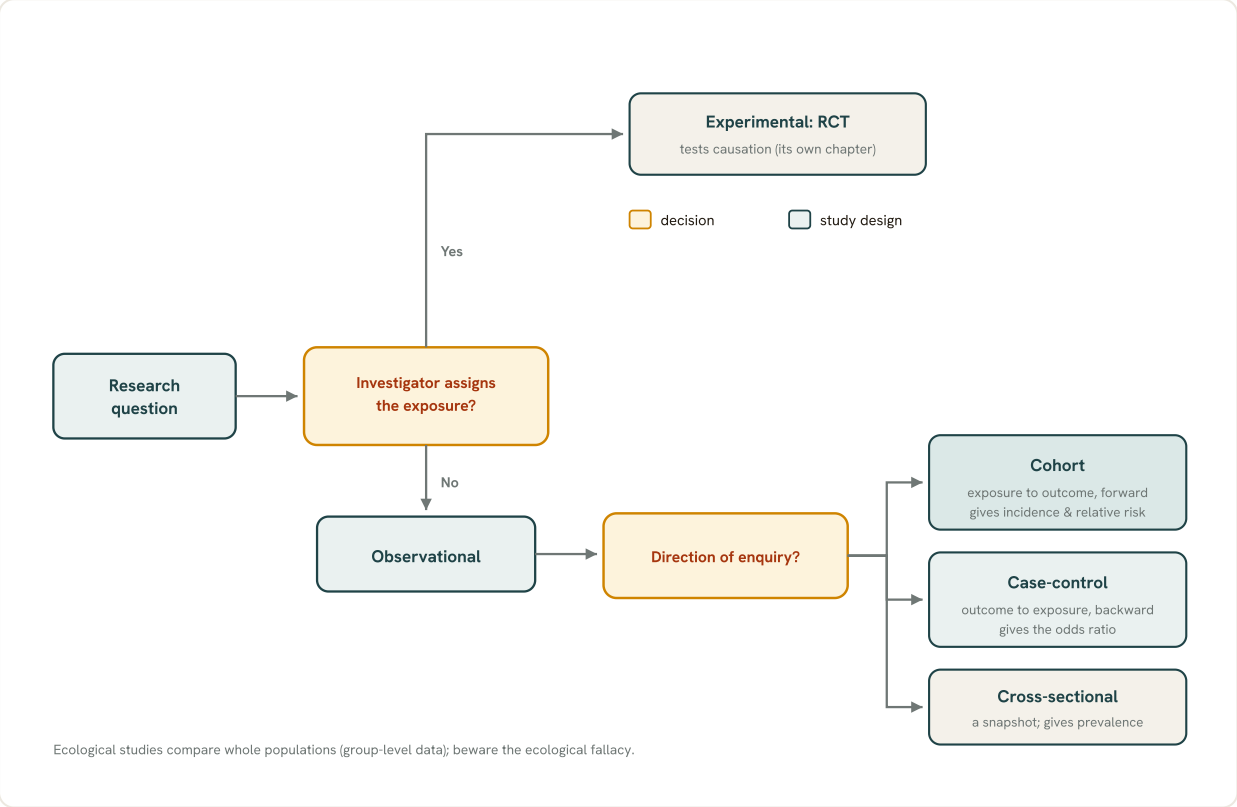



**Fig 7.2** Choosing a study design. The first question is whether the investigator assigns the exposure: if so the design is experimental (the randomised controlled trial, developed in its own chapter). If not, the study is observational, and the direction of enquiry then decides the design: following exposure forward to outcome is a cohort (yielding incidence and the relative risk), tracing outcome back to exposure is a case-control (yielding the odds ratio), and measuring both at once is cross-sectional (yielding prevalence).

### 2.1 Observation versus intervention: the first fork

**Who assigns the exposure?** The primary division of all analytic study designs is **intervention vs observation**. In an interventional (experimental) study, most rigorously the randomised controlled trial, the investigator allocates the exposure or treatment; randomisation balances known and unknown confounders and gives the design its causal strength (Bland 2015). In an **observational** study the investigator only measures what is already happening, so confounding must be handled at analysis (matching, stratification, regression) rather than by design. Everything in this topic lives on the observational side of that fork; the trial designs are treated with the experimental methods. Descriptive observational work (case reports, case series, correlational surveys) describes without a comparison group; the analytic designs below all carry a comparison group and can test associations.

RULE **Name the fork first.** If the investigator assigned the exposure it is a trial; if they only observed it, it is one of the observational designs, and you then read the direction of enquiry.

### 2.2 The cross-sectional study: the snapshot

**Exposure and outcome measured together, at one point in time.** A **cross-sectional** study measures exposure and outcome simultaneously in a defined population at a single point, a snapshot (Bland 2015; Altman 1991). Because everyone is examined once, it is the natural design for a prevalence survey: it counts how many people currently have the condition. It is quick, comparatively cheap, needs no follow-up, and can study several exposures and several outcomes at once, which is why community psychiatric morbidity surveys use it. Its defining weakness is that it *cannot establish temporality*: measuring cause and effect at the same instant, you usually cannot tell which came first, so it is weak for causal inference and prone to reverse causation. It also over-represents long-lasting (prevalent) cases and under-represents brief or rapidly fatal ones.

#### Direction of enquiry

None: exposure and outcome are read simultaneously.

#### Measure yielded

Prevalence; and the prevalence ratio as a crude measure of association.

#### Best for

Prevalence surveys, health-service planning, generating hypotheses.

INDIA

The National Mental Health Survey of India (2015 to 2016) is the flagship cross-sectional example: a snapshot across 12 states giving a current weighted prevalence of any mental disorder of about **10.6** percent (lifetime near **13.7** percent) and a treatment gap for common mental disorders exceeding **80** percent. A snapshot design is exactly right for asking "how many people have the disorder now and how many are in treatment", but it cannot tell you what caused the disorder.

### 2.3 Prevalence and incidence: what each design can count

**Existing cases versus new cases.** The measure a design can produce depends on its time axis, so the two frequency measures anchor the whole topic. **Prevalence** is the proportion of a population that *has* the condition at a given time (existing cases); it is what a cross-sectional snapshot yields. **Incidence** is the rate of *new* cases arising over a period in a population initially free of the disease; it requires follow-up over time and so is the property of the cohort design. The link is that prevalence rises with both incidence and duration (roughly, prevalence is proportional to incidence multiplied by average duration in a steady state), which is why chronic disorders look common on a snapshot even when few new cases arise.

TRAP **Prevalence is not incidence.** A cross-sectional survey gives prevalence (existing cases); only a cohort followed over time gives incidence (new cases). Do not report a snapshot as an incidence.

### 2.4 The case-control study: from outcome, looking back

**Start with the outcome, look back for the exposure.** A **case-control** study assembles a group who already have the outcome (cases) and a comparable group who do not (controls), then looks back to compare how often each group was exposed. It is therefore **retrospective** in direction: the enquiry runs from outcome to exposure (Bland 2015; Altman 1991). Because you start from the diseased, you can study a rare outcome efficiently with a modest sample, examine several exposures at once, and get an answer quickly and cheaply. Its measure of association is the odds ratio (the odds of exposure among cases divided by the odds among controls). The design's weaknesses are its Achilles heel in exams: it is vulnerable to recall bias (cases search their memories harder for exposures than controls do) and to selection bias (through the way controls are chosen), and it cannot directly measure incidence or, ordinarily, relative risk.

#### Direction of enquiry

Outcome to exposure (backward): starts from cases and controls.

#### Measure yielded

The odds ratio (OR); OR approximates the relative risk only when the outcome is rare.

#### Best for

Rare outcomes, diseases with long latency, quick and cheap enquiry, multiple exposures.

#### Main weakness

Recall bias and selection bias; cannot give incidence.

---

TRAP **Case-control gives an odds ratio, not a relative risk.** Because there is no denominator population followed over time, a case-control study cannot yield incidence or a true relative risk; the OR only approximates the RR when the disease is rare.

---

### 2.5 The nested case-control study

**A case-control study built inside a cohort.** The **nested case-control** design draws its cases and controls from within an already-assembled cohort: as outcomes accrue in the cohort, each case is matched to one or more members of the same cohort who were at risk at that time but had not (yet) developed the outcome. Because exposure was recorded at cohort entry, before any outcome, the nested design largely defeats recall bias, and because controls come from the same source population as the cases it reduces selection bias. It is efficient when a costly exposure measurement (an assay on stored blood, an expensive biomarker) need only be run on the sampled cases and controls rather than the whole cohort. It keeps the temporal clarity of the cohort while retaining the economy of the case-control approach.

### 2.6 The cohort study: from exposure, following forward

**Start with the exposure, follow forward to the outcome.** A **cohort** study starts by defining a group free of the outcome, classifies them by exposure status, and follows them forward in time to see who develops the outcome (Bland 2015; Altman 1991). Its direction of enquiry runs from exposure to outcome, and when the following is done in real time it is **prospective**. Because it observes new cases arising in a defined population over time, the cohort design is the one that measures **incidence** directly and yields the relative risk (incidence in the exposed divided by

incidence in the unexposed). It is the design of choice for a rare exposure (you deliberately oversample the exposed at entry), it can examine multiple outcomes of a single exposure, and its temporal sequence (exposure demonstrably precedes outcome) makes it the strongest observational design for causal inference. Its costs are the mirror image: large samples, long follow-up, expense, and loss to follow-up, and it is inefficient for a rare outcome.

**Direction of enquiry**

Exposure to outcome (forward): starts from exposed and unexposed.

**Measure yielded**

**Incidence** and the relative risk (RR); the incidence rate ratio.

**Best for**

Rare exposures, multiple outcomes, establishing temporality.

**Main weakness**

Slow, costly, loss to follow-up; inefficient for rare outcomes.

### 2.7 Prospective versus retrospective cohorts

**Same design, different position of the observer in time.** A **cohort** can be run two ways. In a **prospective** cohort the investigator defines the cohort now, measures exposure now, and waits for outcomes to accrue: exposure data are collected before any outcome, which maximises data quality but demands time and money. In a retrospective (historical) cohort the whole exposure-to-outcome sequence has already happened by the time the study begins; the investigator reconstructs it from existing records (occupational registers, case notes, linked health databases), classifying past exposure and then counting outcomes that have already occurred. Both are true cohort studies (they follow exposure forward to outcome and yield incidence and relative risk); the retrospective form is faster and cheaper but is limited by the completeness and quality of the pre-existing records. Do not confuse the retrospective *cohort* (still exposure to outcome) with the retrospective *direction* of a case-control study (outcome to exposure).

**WORKED EXAMPLE**

Read a 2x2 table by design. In a **cohort** of 200 exposed and 200 unexposed, followed forward, 40 exposed and 10 unexposed develop the disorder. Incidence in exposed = 40/200 = **0.20**; incidence in unexposed = 10/200 = **0.05**; relative risk = 0.20 / 0.05 = **4.0** (the exposed are four times as likely to develop the outcome). Now take the same table as a **case-control** study (50 cases, 350 controls): odds of exposure in cases = 40/10 = 4; odds of exposure in controls = 160/190; odds ratio = (40 x 190) / (10 x 160) = 7600 / 1600 = **4.75**. Same underlying data, read in two directions: the cohort reads *across* the exposure rows (incidence in the exposed is a / (a + b)) to give an RR, the case-control reads *down* the disease columns (fixing the outcome, then comparing exposure odds) to give an OR, and the OR overshoots the RR because the outcome here is not rare.

### 2.8 The ecological study and the ecological fallacy

**The unit of analysis is the group, not the person.** An **ecological** study correlates an exposure and an outcome measured at the level of populations (countries, states, districts, time periods) rather than individuals: for example, plotting national fish consumption against national

depression rates, or state lithium-in-water levels against state suicide rates. It is cheap and fast because it reuses routinely collected aggregate data, and it is a useful hypothesis-generating and hypothesis-testing tool at the population level. Its signature hazard is the ecological fallacy: an association that holds between group averages need not hold for the individuals within those groups, so a group-level correlation cannot be assumed to describe individual-level risk. It is the price of losing the individual as the unit of analysis, and it is the single most examined criticism of the ecological design.

TRAP **The ecological fallacy.** Inferring an individual-level relationship from a group-level (aggregate) correlation is invalid; a correlation between country averages tells you nothing certain about any individual.

### 2.9 The design decision logic: intervention, direction, time axis

**Three questions choose the design.** Faced with a vignette, decide in order. First, *intervention vs observation*: did the investigator assign the exposure (trial) or merely observe it (one of these designs)? Second, the *direction of enquiry*: does the study start from the outcome and look back (case-control) or start from the exposure and look forward (cohort), or read both at once (cross-sectional)? Third, the *time axis*: a single snapshot (cross-sectional, gives prevalence), backward from outcome (case-control, gives an odds ratio), or forward from exposure (cohort, gives incidence and relative risk). Layered onto that is the frequency logic that decides between the two comparative designs: match the rarity of the exposure or the outcome to the design that captures it efficiently.

| DESIGN | DIRECTION OF ENQUIRY | MEASURE OF ASSOCIATION | BEST FOR | MAIN WEAKNESS |
|---|---|---|---|---|
| **Cross-sectional** | Simultaneous (snapshot) | **Prevalence (prevalence ratio)** | Prevalence surveys; service planning; several exposures and outcomes at once | No temporality; reverse causation; favours long-duration cases |
| **Case-control** | Outcome to exposure (backward, retrospective) | **Odds ratio (OR)** | Rare outcomes; long-latency disease; quick and cheap; multiple exposures | Recall bias; selection bias; no incidence, no true RR |
| **Nested case-control** | Outcome to exposure, within a cohort | **Odds ratio (OR)** | Costly exposure assays; when exposure was recorded before outcome | Still smaller than the full cohort; needs the parent cohort |
| **Cohort** | Exposure to outcome (forward, prospective) | **Relative risk (RR); incidence** | Rare exposures; multiple outcomes; establishing temporality | Slow, costly, loss to follow-up; poor for rare outcomes |

| DESIGN | DIRECTION OF ENQUIRY | MEASURE OF ASSOCIATION | BEST FOR | MAIN WEAKNESS |
|---|---|---|---|---|
| **Ecological** | Group-level correlation | **Correlation of population rates** | Cheap hypothesis generation from aggregate data | Ecological fallacy: group data cannot be read to the individual |

The observational family read across design, direction, measure, best use and main weakness; the hit column is the measure of association each design yields.

### 2.10 The two selectors: choosing a design, choosing a test

**Match the question to the design; match the data to the test.** Two interleaved selectors carry a research vignette from question to result. The first turns a clinical question into a design; the second turns the resulting data into the correct statistical test. The design chooser turns on the rarity of what is being studied and on whether temporality matters; the test chooser turns on the type of data, the number of groups, and whether the groups are paired or unpaired.

| QUESTION / SITUATION | DESIGN TO PICK |
|---|---|
| **How common is the condition now?** | **Cross-sectional (prevalence)** |
| **Rare outcome, or a long-latency disease?** | **Case-control** |
| **Rare exposure, or many outcomes of one exposure?** | **Cohort** |
| **Need to prove exposure preceded outcome (temporality)?** | **Prospective cohort** |
| **Costly exposure assay, cases arising inside a cohort?** | **Nested case-control** |
| **Only aggregate, population-level data available?** | **Ecological (mind the fallacy)** |

Study-design chooser: read the clinical question, pick the design (question to design).

| DATA TYPE | GROUPS / PAIRING | CORRECT TEST |
|---|---|---|
| **Continuous, normal** | 2 unpaired groups | **Unpaired (independent) t-test** |
| **Continuous, normal** | 2 paired groups | **Paired t-test** |
| **Continuous, normal** | 3+ groups | **One-way ANOVA** |
| **Continuous, non-normal** | 2 unpaired groups | **Mann-Whitney U** |
| **Continuous, non-normal** | 2 paired groups | **Wilcoxon signed-rank** |
| **Continuous, non-normal** | 3+ groups | **Kruskal-Wallis** |
| **Categorical (counts)** | 2+ unpaired groups | **Chi-squared test** |
| **Categorical, paired** | 2 paired groups | **McNemar test** |

Statistical-test chooser: data type by number of groups by pairing, giving the correct parametric or non-parametric test.

**HOLD THIS**

Case-control starts from the outcome and gives an odds ratio; cohort starts from the exposure and gives incidence and a relative risk; cross-sectional is a snapshot and gives prevalence.

**COMMON TRAP**

Two mirror-image errors examiners set. First, reporting a **relative risk** from a case-control study, which cannot yield one (it gives an odds ratio; the OR approximates RR only when the outcome is rare). Second, choosing a case-control design for a rare *exposure* (it is efficient for a rare *outcome*): a rare exposure calls for a cohort that oversamples the exposed. Keep the pairing straight: rare outcome to case-control, rare exposure to cohort.

**PEARL**

The odds ratio and the relative risk converge when the outcome is rare (the "rare disease assumption"): with a low outcome frequency the OR from a case-control study is a good approximation of the RR you would have got from a cohort. As the outcome becomes common the OR increasingly exaggerates the RR, which is exactly why a case-control OR should never be quoted as a relative risk when the outcome is frequent.

**MNEMONIC**

**"Cohort = Cause first; Case-control = Consequence first."**

The cohort begins with the putative cause (the exposure) and follows forward; the case-control begins with the consequence (the outcome) and looks back. Both start with C, so anchor on the second word: cause-first is forward-looking cohort (relative risk), consequence-first is backward-looking case-control (odds ratio).

**GOOD TO REMEMBER**

- **Cross-sectional:** measures exposure and outcome at once, a snapshot; use it for a prevalence survey; gives prevalence and the prevalence ratio, and cannot establish temporality.
- **Case-control:** starts from the outcome and looks back for exposure (retrospective); use it for a rare outcome or long-latency disease; gives the odds ratio, OR = (odds of exposure in cases) / (odds of exposure in controls).
- **Nested case-control:** a case-control study drawn from within a cohort; use it when a costly exposure assay makes measuring the whole cohort wasteful; gives an odds ratio while defeating recall bias.
- **Cohort:** starts from exposure and follows forward to outcome (prospective); use it for a rare exposure or multiple outcomes; gives incidence and the relative risk, RR = (incidence in exposed) / (incidence in unexposed).
- **Ecological:** correlates exposure and outcome at the group level; use it for cheap hypothesis generation from aggregate data; the number to fear is the ecological fallacy, that group correlations do not transfer to individuals.

## 3 · The randomised controlled trial and drug-trial phases

The randomised controlled trial (RCT) is the only truly **experimental design** in clinical research and the acknowledged gold standard for testing whether a treatment actually works (Companion to Psychiatric Studies 2010). Everything else in this chapter observes; the RCT intervenes, allocates by chance, and so earns the right to speak about cause. The examiner rarely asks "what is an RCT"; the marks sit in the machinery that protects it: randomisation, allocation concealment, blinding, and an analysis that keeps the randomised groups intact.

Read this topic as a chain in which each link defends the last. Randomisation creates comparable groups; allocation concealment stops the recruiter from breaking the randomisation at entry; blinding stops patients, clinicians and assessors from biasing what happens afterwards; intention-to-treat keeps the analysis faithful to the randomisation. Break any link and the design degrades toward a mere observational comparison.

### 3.1 The randomised controlled trial as the experimental design

**Only a controlled trial with random allocation is truly experimental.** In a **randomised controlled** trial the investigator assigns each consenting participant to an intervention or a comparator by a chance mechanism, follows both arms forward, and compares a pre-specified outcome. Because the investigator controls the exposure and allocates it at random, the RCT (equivalently the **randomized controlled** trial in American spelling) is the design least vulnerable to selection bias and confounding, which is why it anchors the top of the evidence hierarchy for questions of therapy. Non-randomised comparisons tend to over-estimate benefit because the newer treatment is quietly given to less severe or better-prognosis patients, the problem of confounding by indication. Even within randomised trials, inadequate or unclear allocation concealment inflates the estimated treatment effect, by around **30% to 41%** on average (Schulz et al 1995).

### 3.2 Randomisation: why it works, and its varieties

**Randomisation balances known and, crucially, unknown confounders.** Randomisation (American: randomization) has two purposes. First and foremost it distributes both the measured prognostic factors (age, sex, severity) and, more importantly, the *unmeasured and unknown* ones evenly between arms, so that any difference in outcome can be attributed to the treatment. Second, it removes the recruiter's ability to steer sicker or healthier patients into a preferred arm. No other design can balance confounders you have not even thought to measure, and that is the single reason the RCT out-ranks the cohort study for causal claims about therapy (Bland 2015).

#### Simple randomisation

Each patient allocated independently by chance (the equivalent of a fair coin or a random-number list); can leave unequal group sizes in small trials, but self-corrects as the trial grows large.

#### Block randomisation

Patients randomised in small blocks (of four, six, and so on) so that the numbers in each arm stay equal throughout recruitment; the cost is that the last allocation in a block can become guessable.

#### Stratified randomisation

A separate randomisation schedule is run within each level of a key prognostic variable (for example centre, or baseline severity) so that variable is balanced by design rather than left to chance; usually combined with blocks.

RULE **Random allocation is the only tool that balances unknown confounders.** That is why the RCT, not the cohort study, is the design for a causal therapy question.

### 3.3 Allocation concealment: guarding the moment of assignment

**Concealment hides the *upcoming* assignment from whoever recruits.** Allocation concealment means that the person enrolling a patient cannot know, at the moment of consent, which arm that patient is about to enter. If the schedule is predictable (allocation by date of birth, alternate days, an unsealed list), a recruiter can consciously or unconsciously include or exclude patients to load one arm, and the randomisation is subverted before it happens. The robust solution is a remote or independent system: the recruiter phones or logs in, enters the patient's baseline data, and only then is the allocation released. Trials in which allocation was easy to subvert were far more likely to report a significant effect (trials with subvertible allocation were consistently more likely to report a significant effect; Chalmers et al 1983).

TRAP **Allocation concealment and blinding are not the same thing.** Concealment protects the assignment *before* and *at* entry; blinding protects perception *after* entry. A trial can be double-blind yet have had broken concealment, and vice versa.

### 3.4 Blinding: what single, double and triple each conceal

**Blinding stops knowledge of the assigned arm from biasing what comes next.** Where allocation concealment guards entry, blinding (masking) guards everything after it by keeping the arm hidden from one or more parties. A **placebo** is the instrument that makes patient-level blinding possible, an inert comparator indistinguishable from the active drug. The purpose of single, double and **double-blind** or triple blinding is to reduce observer, performance and ascertainment bias.

| LEVEL | WHO IS KEPT UNAWARE OF THE ARM | BIAS IT CHIEFLY PREVENTS |
|---|---|---|
| **Open (unblinded)** | Nobody; all parties know the arm | None; maximally biased |
| **Single-blind** | **The patient only** | Patient expectation / placebo response |
| **Double-blind** | **Patient and treating clinician** | Patient expectation plus clinician performance bias |
| **Triple-blind** | **Patient, clinician and the outcome assessor / analyst** | Adds ascertainment (assessment) bias |

Each blinding level conceals the arm from one more party; the extra party in triple-blind is the outcome assessor or data analyst (Companion to Psychiatric Studies 2010).

**Blinding is often imperfect.** Distinctive side-effects (anticholinergic dryness, extrapyramidal signs) can unmask the active drug; a useful check is to ask patients and raters to guess their arm

at the end, where successful blinding gives no-better-than-chance guessing (Even et al 2000).

### 3.5 Intention-to-treat versus per-protocol analysis

**Analyse everyone in the arm to which they were randomised.** Intention-to-treat (ITT) analysis includes every randomised participant in their originally allocated arm, regardless of whether they completed, complied with, or even received the treatment. Per-protocol (completer) analysis restricts to those who finished per protocol. ITT is preferred because the moment you exclude dropouts or non-compliers you re-introduce exactly the selection you randomised away: dropout is rarely random (people leave because a drug failed or was not tolerated), so a completer analysis flatters the treatment. ITT preserves the randomisation and therefore keeps the groups comparable and the comparison unbiased (Altman 1991). It gives the more conservative, more clinically realistic estimate of effectiveness; per-protocol answers the narrower "did it work in those who took it" question and is more prone to bias.

RULE **ITT analyses everyone as randomised.** "Once randomised, always analysed" preserves the balance that randomisation bought and gives the honest, conservative effect estimate.

For missing outcomes, last-observation-carried-forward (LOCF) has been used to complete ITT datasets but is statistically dubious, since trial dropout in psychiatry is often substantial and treatment-related, so the missing values are rarely missing at random (Companion to Psychiatric Studies 2010).

### 3.6 Trial architectures: parallel-group, crossover, factorial, cluster randomised

**The architecture is chosen to fit the disease and the question.** The default is the parallel-group trial, in which independent arms run side by side. Three important variants exist. A crossover trial gives every participant each treatment in turn in random order, so each patient serves as their own control. A factorial design tests two or more interventions simultaneously in one trial (a two-by-two factorial gives A, B, both, or neither), efficiently answering more than one question and probing interaction. A cluster randomised (American: cluster randomized) trial randomises whole groups, such as clinics or general-practice teams, rather than individuals.

| ARCHITECTURE | STRUCTURE | WHEN TO USE IT | THE CATCH |
|---|---|---|---|
| **Parallel-group** | Separate arms compared concurrently | **The default for most therapeutic questions** | Needs the full sample size in each arm |
| **Crossover** | **Each patient receives both treatments in random sequence, own control** | Rare or stable chronic disease where recruiting large parallel groups is impractical | Carryover and order effects; needs a washout; not for curable or unstable conditions |
| **Factorial** | **Two-plus interventions crossed in one trial (2x2: A, B, both, neither)** | Two questions, or a suspected interaction, tested economically at once | Assumes no strong interaction between the factors; complex to interpret if there is one |

| ARCHITECTURE | STRUCTURE | WHEN TO USE IT | THE CATCH |
|---|---|---|---|
| **Cluster randomised** | **Groups (clinics, practices, wards) randomised, not individuals** | Service-level or educational interventions that cannot be given to one person without contaminating neighbours | The cluster is the unit of analysis; needs many clusters for power |

Crossover buys power in rare stable diseases at the price of carryover; factorial buys a second answer if the factors do not interact; cluster is forced whenever the intervention cannot be confined to an individual (Companion to Psychiatric Studies 2010).

### 3.7 Superiority, non-inferiority and equivalence framing

**The objective determines the hypothesis and the sample size.** A **superiority** trial asks whether the new treatment beats the comparator (or placebo). A **non-inferiority** trial asks whether the new treatment is not worse than an established one by more than a pre-specified acceptable margin, used when a new agent offers a side benefit (cheaper, safer, easier) and a placebo arm would be unethical. An **equivalence** trial asks whether two treatments differ by no more than a margin in either direction, typical for bioequivalence. The direction of the question changes the null hypothesis, so a non-inferiority or equivalence claim cannot simply be read off a failed superiority test.

**COMMON TRAP**

"No significant difference" in a superiority trial does not prove the treatments are equivalent; absence of evidence is not evidence of absence. Equivalence and non-inferiority must be designed and powered as such from the outset, with a pre-stated margin. A small, underpowered superiority trial that finds nothing is uninformative, not reassuring.

### 3.8 The drug trial phases

**A new drug is licensed by climbing four phases.** Before an RCT can even ask about efficacy, a compound must pass through the staged programme of a drug trial. Each phase has a distinct question and a characteristically larger sample. **Phase i** is first-in-human: small groups of usually healthy volunteers receive the drug to establish safety, tolerability and pharmacokinetics (in oncology and some psychiatry contexts, patients rather than healthy volunteers are used). **Phase ii** moves to patients with the target disorder to look for preliminary efficacy and to find the dose. **Phase iii** is the large, comparative, confirmatory RCT that supplies the pivotal evidence for licensing. Phase IV is post-marketing surveillance after approval, detecting rare or long-term adverse effects in real-world use (Goodman & Gilman 2018).

| 20 to 100 | 100 to 300 | 300 to 3000 | thousands |
|---|---|---|---|
| PHASE I (HEALTHY VOLUNTEERS) | PHASE II (PATIENTS) | PHASE III (PATIENTS) | PHASE IV (POST-MARKETING) |

| PHASE | CORE QUESTION | WHO / TYPICAL N | DESIGN |
|---|---|---|---|
| **Phase I** | **Is it safe in humans, and how is it handled (PK)?** | Small healthy volunteer groups, roughly **20 to 100** | First-in-human, dose-escalation, often open |
| **Phase II** | **Does it show efficacy, and at what dose?** | Patients with the disorder, roughly **100 to 300** | Dose-finding; often placebo-controlled, randomised |
| **Phase III** | **Does it work better than the comparator, confirmed?** | Larger patient samples, roughly **300 to 3000** | Large double-blind confirmatory RCT for licensing |
| **Phase IV** | **What happens in real-world, long-term use?** | Post-marketing, thousands of users | Surveillance / pharmacovigilance after approval |

The question, the sample and the setting all scale up phase by phase: safety in a handful, efficacy and dose in hundreds, confirmation in thousands, then lifelong surveillance (typical Ns per standard pharmacology texts; Goodman & Gilman 2018).

**WORKED EXAMPLE**

A novel antidepressant: phase I gives single ascending oral doses to about 40 healthy volunteers, measuring half-life and tolerability with no efficacy claim. Phase II randomises roughly 200 patients with major depression across three doses versus placebo to identify the effective dose. Phase III runs two double-blind placebo-controlled RCTs of around 500 patients each, the pivotal evidence submitted for licensing. After approval, phase IV registries pick up a rare hepatic signal invisible in the pre-licensing sample. The same drug thus needs every phase; a positive phase III cannot substitute for phase IV vigilance.

### 3.9 How the pieces defend one another

**The safeguards form one continuous chain.** Randomisation manufactures comparable groups; allocation concealment stops that randomisation being unpicked at recruitment; blinding stops it being eroded during treatment and assessment; intention-to-treat stops it being discarded at analysis. Assess any published RCT by walking this chain, and you will locate its weakest link. A trial that is randomised but with poor concealment, or double-blind but analysed per-protocol, has a leak precisely where the exam wants you to point.

**HOLD THIS**

Allocation concealment guards the assignment at entry; blinding guards perception after entry; intention-to-treat keeps everyone analysed in the arm they were randomised to; and the drug phases climb safety (20 to 100 healthy volunteers) to efficacy/dose (100 to 300 patients) to confirmatory licensing (300 to 3000) to post-marketing surveillance.

**GOOD TO REMEMBER**

- **Randomised controlled trial:** the only experimental therapeutic design; allocates treatment by chance and compares arms forward; the one design that balances unknown confounders, which is why it is the therapy gold standard.
- **Randomisation:** distributes known and unknown confounders evenly; simple (self-corrects when large), block (keeps arms equal), stratified (balances a key prognostic factor by design).
- **Allocation concealment:** hides the upcoming assignment from the recruiter at the moment of entry; best achieved by remote / independent randomisation; distinct from blinding, which acts after entry.
- **Blinding:** conceals the arm to cut bias; single (patient), double (patient plus clinician), triple (plus outcome assessor); a placebo makes patient blinding possible.
- **Intention-to-treat vs per-protocol:** ITT analyses all as randomised and preserves comparability (conservative); per-protocol counts only completers and is prone to bias; "once randomised, always analysed".
- **Parallel-group:** concurrent independent arms; the default; needs full N per arm.
- **Crossover:** each patient is own control receiving both treatments in sequence; use in rare, stable, chronic disease; watch carryover and order effects, use a washout.
- **Factorial:** two-plus interventions crossed (2x2 = A, B, both, neither); answers two questions and tests interaction in one trial; valid if the factors do not strongly interact.
- **Cluster randomised:** randomises groups (clinics, practices) not individuals; use for service or educational interventions that cannot be confined to one person; needs many clusters because the cluster is the analysis unit.
- **Superiority / non-inferiority / equivalence:** beats comparator / not worse than by a set margin / differs by no more than a margin either way; a non-significant superiority result does not prove equivalence.
- **Drug trial phases:** phase I safety and PK in 20 to 100 healthy volunteers; phase II efficacy and dose-finding in 100 to 300 patients; phase III large confirmatory comparative RCT of 300 to 3000 for licensing; phase IV post-marketing surveillance in thousands.

## 4 · Validity, bias and confounding

**This is the quality-control chapter of research.** Every study makes a claim; validity, bias and confounding are the three questions the examiner uses to decide whether the claim is believable. **Validity** asks whether the finding is true (for this sample, and then for the wider world); **bias** is systematic error that pushes the answer consistently in one direction; and **confounding** is a third variable that mimics a real association. The stem usually gives a scenario and asks you to name the flaw, classify it, and say how it should have been controlled. Get the taxonomy clean and almost every methods question falls open (Kaplan & Sadock 2024; Bland 2015; Altman 1991).

Two housekeeping distinctions anchor the topic before the detail. First, keep **systematic error** (bias, which no increase in sample size removes) apart from **random error** (chance, which shrinks as the sample grows and is quantified by confidence intervals and the p-value). Second, the study reliability and validity discussed here (are the study's results reproducible and true) are a DIFFERENT idea from the psychometric-test reliability and validity of a scale or instrument (whether a questionnaire measures consistently and measures what it claims); those are taught in the Psychometry, Assessment and Descriptive Psychopathology notes. Keep the two vocabularies

separate: here the object is a whole study, there it is a single test; this separation is restated where reliability is defined below.

### 4.1 Internal validity: are the results true for this sample

**Does the study answer its own question correctly, free of bias and confounding?** Internal validity is the degree to which the observed association between exposure and outcome is real for the participants actually studied, rather than an artefact of bias, confounding or chance (Kaplan & Sadock 2024). It is the first and non-negotiable property: a study with poor internal validity is simply wrong, and how far it generalises is then irrelevant. Threats to internal validity are exactly the topics of this chapter, selection bias, information bias and confounding, plus chance. The randomised controlled trial earns its rank because randomisation and blinding protect internal validity better than any observational design can. Efficacy trials in tightly selected samples are built to maximise internal validity (Kaplan & Sadock 2024).

RULE **Internal validity comes first.** If the result is not true for the sample studied, it does not matter whether it generalises: settle internal validity before external validity.

### 4.2 External validity: do the results generalise

**Can the finding be applied beyond the study sample?** External validity, also called **generalisability** (spelled generalisability or generalizability), is the extent to which the results hold for patients outside the study, in other settings, populations and times (New Oxford Textbook of Psychiatry). A trial can have flawless internal validity yet poor external validity if its participants were so narrowly selected (single-diagnosis, no comorbidity, no substance use) that the result cannot be read across to the messy patients of ordinary practice. This is the efficacy versus effectiveness trade-off: efficacy studies maximise internal validity in narrow samples, whereas effectiveness (pragmatic) studies use broad, typical, comorbid populations so that external validity carries more weight (Kaplan & Sadock 2024). The two validities pull against each other, and reading which one a design has bought is a favourite question.

WORKED EXAMPLE

An antidepressant RCT enrols only patients aged 18 to 45 with major depression and no comorbidity, randomises and double-blinds cleanly, and reports a clear drug effect. Its internal validity is high: the effect is real for those studied. Its external validity is low: most real depressed patients are older, or have anxiety, alcohol use or physical illness, so the result may not generalise to them. High internal, low external validity is the signature of a tightly controlled efficacy trial.

### 4.3 Reliability: reproducibility of the measurement

**Does the study measure the same thing consistently on repetition?** Study reliability, or **repeatability**, is the reproducibility of a measurement: if the same subjects were measured again under the same conditions, would the same values result. Reliability concerns precision (random scatter), whereas validity concerns accuracy (systematic closeness to the truth); the classic image is the target, where a reliable-but-invalid instrument lands its shots tightly clustered but off the bullseye, and a valid-but-unreliable one scatters around the centre. A measurement can be reliable

without being valid, but it cannot be valid without first being reliable, so reliability is necessary but not sufficient for validity. As flagged in the intro, this study-level reliability is distinct from the psychometric-test reliability (internal consistency, test-retest, inter-rater) covered in the Psychometry, Assessment and Descriptive Psychopathology notes.

### 4.4 Bias: systematic error, and its two great families

**Bias is systematic error, built into the study, that no larger sample can fix.** Bias is any systematic (non-random) error in the design, conduct or analysis of a study that produces a result consistently distorted away from the truth (Kaplan & Sadock 2024). Because it is systematic, adding participants only makes a biased study more precisely wrong; bias is defeated by better design, not by size. Following the standard epidemiological division, bias falls into two great families plus a literature-level form: selection bias (who gets into the study and who is compared), information bias (how exposure and outcome are measured once in), and publication bias (which studies reach print). Naming the family first, then the specific mechanism, is the reliable way to answer a bias vignette (Tasman Psychiatry).

### 4.5 Selection bias

**A systematic error in how participants enter or are compared.** Selection bias arises when the way subjects are selected, or the way exposed and unexposed (or cases and controls) are chosen, is related to both the exposure and the outcome, so the compared groups differ systematically from the start (Tasman Psychiatry). Textbook varieties include the **healthy worker effect** (employed populations are healthier than the general population, so occupational cohorts understate harm), **self-selection (volunteer) bias** (people who respond to advertisements or agree to take part differ from those who do not, and the exposed with symptoms may be keener to enrol), **Berkson's bias** (hospital-based cases and controls carry a spurious association because admission itself depends on having one or more conditions), and **loss to follow-up (attrition) bias** (differential dropout in a cohort or trial, a serious threat when dropout is related to both exposure and outcome). The defence is at the design stage: define the source population, sample and recruit exposed and unexposed identically, and choose controls from the same population that produced the cases.

### 4.6 Information bias, including recall and observer bias

**A systematic error in how exposure or outcome is measured.** Information bias (also called observation or measurement bias) occurs when data on exposure or outcome are obtained in a non-comparable manner between the groups, so that one group's information is systematically more or less accurate than the other's (Tasman Psychiatry). Its named forms are high-yield. Recall bias is differential accuracy of memory: people with the outcome (cases, or mothers of an affected child) search their past for exposures more thoroughly than unaffected controls, inflating the apparent association; it is the classic weakness of the retrospective case-control study. **Observer (or interviewer) bias** is systematic error introduced by the assessor: an interviewer who knows a subject's disease status probes exposure harder in cases, or an unblinded rater scores the outcome more favourably in the treated arm; masking (keeping the assessor blind to exposure or allocation) is the remedy (Tasman Psychiatry). **Misclassification** is the mechanism by which

information bias operates: non-differential misclassification (errors unrelated to group) usually dilutes an association toward the null, whereas differential misclassification (errors that differ by group, as in recall bias) can bias in either direction and is the more dangerous. Publication bias is a special, literature-level information problem addressed at 4.7.

TRAP **Recall bias needs two groups remembering differently.** It is not just "poor memory": it is differential recall between cases and controls. Symmetric forgetting in both groups is non-differential misclassification, not recall bias.

### 4.7 Publication bias

**Positive studies are published more, and sooner, than negative ones.** Publication bias is the tendency for studies with statistically significant, positive or striking results to be submitted, accepted and published more readily than studies with null or negative findings, so the published literature systematically overstates an effect. It is the reason a meta-analysis restricted to published trials can be misleading, and it is detected within the systematic-review and meta-analysis topic by an asymmetric funnel plot (and formal tests such as Egger's), and mitigated by trial registries and grey-literature searching. It is treated more fully as a threat to the systematic review; here, hold that it is a bias of the literature, not of a single study's internal conduct.

### 4.8 Confounding: the third variable

**A third variable linked to both exposure and outcome, and not on the causal path.** Confounding occurs when a third factor is associated with the exposure and is independently a cause of the outcome, so that it distorts (creates, inflates, masks or even reverses) the apparent exposure to outcome association (Tasman Psychiatry). Three conditions define a **confounder**: it must be associated with the exposure, it must be an independent risk factor for (cause of) the outcome, and it must NOT lie on the causal pathway between exposure and outcome (a variable on that pathway is a mediator, not a confounder). The examiner's classic is the unemployment to suicidal behaviour question in which **alcoholism** is related both to unemployment and to suicidal behaviour, and therefore confounds the association unless it is adjusted for (Tasman Psychiatry). Other stock examples are age and sex confounding almost any exposure to disease link, and the coffee to lung-cancer association confounded by smoking (smokers drink more coffee and smoking causes lung cancer). Spot a confounder by testing the three conditions, then decide whether it was handled at design or at analysis.

WORKED EXAMPLE

A study finds unemployed people have more suicidal behaviour, crude relative risk **2.0**. But alcohol dependence is commoner among the unemployed AND independently raises suicide risk, so it satisfies both confounder arms and is not on the causal path. Stratify by alcohol status: within non-drinkers RR is **1.2** and within dependent drinkers RR is **1.3**. The crude 2.0 was inflated by confounding; the adjusted (stratum-specific) estimate near 1.2 to 1.3 is the truer effect of unemployment itself. Adjusting for alcohol is correct here because alcohol is a confounder, not a step on the pathway from unemployment to suicide.

### 4.9 Controlling confounding at the design stage

**Prevent the imbalance before data are collected.** Four design tools handle confounding up front. Randomisation allocates subjects to groups by chance so that, on average, both known AND unknown confounders are balanced between arms; it is the only method that controls confounders you have not measured or even thought of, which is the entire reason the RCT is the reference standard. **Restriction** admits only one level of the confounder (for example, study only non-smokers), which removes its confounding effect but narrows external validity. **Matching** selects comparison subjects to have the same distribution of a confounder as the index group (individual or frequency matching), balancing that specific confounder, though matched data then require a matched analysis. Two further design safeguards protect randomisation itself: allocation concealment hides the upcoming assignment from the recruiter so that selection into arms cannot be manipulated, and blinding (masking) keeps participants, clinicians and assessors unaware of allocation; a **double-blind** design masks both the participant and the assessor, defending against performance and observer (ascertainment) bias. Note that allocation concealment governs entry into the arms whereas blinding governs everything after; they are distinct and both matter.

RULE **Only randomisation controls unknown confounders.** Restriction, matching, stratification, standardisation and regression each control confounders you have measured; solely randomisation balances the ones you have not.

### 4.10 Controlling confounding at the analysis stage

**Correct for the measured confounder after data are collected.** When a confounder was recorded but not designed out, three analytic tools adjust for it. Stratification divides the data into strata of the confounder, computes the association within each stratum where the confounder is held constant, and pools the stratum-specific estimates (the Mantel-Haenszel method); comparing the crude and adjusted estimates reveals how much confounding was present. **Standardisation** re-weights rates to a common reference population (direct or indirect), the method behind age-standardised and sex-standardised rates, so groups with different age or sex structures can be compared fairly. Multivariable (regression) adjustment models the outcome on the exposure plus the confounders simultaneously (linear, logistic or Cox regression), yielding an exposure effect adjusted for every covariate in the model; it is the workhorse of modern observational analysis. The hard limit of every analytic method is that you can only adjust for confounders you measured, and residual confounding from unmeasured or mismeasured variables always remains, which is precisely why randomisation is unique.

TRAP **Never adjust for a mediator.** Confounding is NOT a type of bias, and a variable on the causal pathway (a mediator) is not a confounder: adjusting for it wrongly blocks part of the real effect (overadjustment).

| CLASS | NAMED FORM | WHAT GOES WRONG | TYPICAL DESIGN | DEFENCE |
|---|---|---|---|---|
| **Selection bias** | **Healthy worker effect** | Employed sample healthier than general population, harm understated | Occupational cohort | Suitable external comparison group |
| **Selection bias** | **Self-selection (volunteer)** | Participants differ systematically from non-participants | Advertised recruitment, surveys | Define and sample a source population |
| **Selection bias** | **Berkson's bias** | Hospital cases and controls carry a spurious association via admission | Hospital case-control | Community (population) controls |
| **Selection bias** | **Loss to follow-up (attrition)** | Differential dropout related to exposure and outcome | Cohort, RCT | Minimise and analyse dropout; intention-to-treat |
| **Information bias** | **Recall bias** | Cases remember exposures more thoroughly than controls | Retrospective case-control | Records-based exposure; nested design |
| **Information bias** | **Observer / interviewer bias** | Unblinded assessor probes or scores groups differently | Any assessed study | Masking (blinding) of assessors |
| **Information bias** | **Misclassification** | Errors in exposure/outcome labelling (non-differential dilutes; differential either way) | Any measurement | Validated, objective, blinded measures |
| **Publication bias** | **Positive-result bias** | Significant studies published more, literature overstates effect | Meta-analysis of published trials | Trial registries; funnel plot; grey literature |
| **Confounding** | **Third-variable distortion** | Factor linked to exposure and outcome, not on causal path, mimics association | Any observational study | Randomisation, restriction, matching; stratification, standardisation, regression |

Bias taxonomy: the family, a named form, the mechanism, the design it typically strikes, and the defence; the hit column is the named entity to recognise in a stem.

**HOLD THIS**

Bias is systematic error split into selection bias and information bias; confounding is a third variable tied to both exposure and outcome but not on the causal path; and randomisation is the only tool that balances UNKNOWN confounders.

**COMMON TRAP**

Two errors examiners love. First, calling confounding a type of bias: it is a separate category (a real third-variable effect, controllable at analysis), not a systematic measurement or selection error. Second, treating internal and external validity as the same axis: a narrow efficacy trial can have excellent internal validity and poor external validity at once, and confusing the two mis-reads the whole study.

**PEARL**

To separate confounding from mediation, ask where the third variable sits. If it independently causes the outcome and merely travels with the exposure, it is a confounder, so adjust for it. If the exposure causes it and it in turn causes the outcome, it is a mediator on the causal path, so adjusting for it (overadjustment) erases part of the true effect. The same statistical adjustment is right for the first and wrong for the second.

**MNEMONIC**

**Confounder = "linked to Exposure, causes Outcome, off the Path" (E-O-P).**

A confounder must be associated with the Exposure, be an independent cause of the Outcome, and lie off the causal Path. Fail any one of the three and it is not a confounder: variables on the path are mediators, and variables unrelated to the exposure are simply other risk factors.

**GOOD TO REMEMBER**

- **Internal validity:** are the results true for this sample; threatened by selection bias, information bias, confounding and chance; the one to secure first, before generalisability.
- **External validity (generalisability):** do the results apply beyond the sample; use it to judge whether an efficacy-trial finding transfers to ordinary patients; bought by broad, typical, pragmatic samples.
- **Reliability (repeatability):** reproducibility of a measurement on repetition; the target-clustering property; necessary but not sufficient for validity (reliable can be invalid, but valid needs reliable).
- **Bias:** systematic (not random) error; use design, not sample size, to defeat it; the number to fear is that no increase in n reduces bias.
- **Selection bias:** non-comparable entry or comparison groups; suspect it in volunteer, hospital-based or high-dropout studies; defended by sampling a defined source population identically.
- **Information bias:** non-comparable measurement of exposure or outcome; includes recall bias and observer bias; defended by blinding (masking) and validated measures.
- **Recall bias:** differential memory between cases and controls; the classic flaw of the retrospective case-control study; defended by records-based or nested designs.
- **Confounding:** a third variable associated with exposure and causing the outcome, off the causal path; controlled at design (randomisation, restriction, matching, allocation concealment, blinding) or at analysis (stratification, standardisation, regression); only randomisation handles unknown confounders.

## 5 · Sampling methods

**Whether a study can speak beyond its own participants is decided the moment the sample is drawn.** **Sampling** is the process of drawing a subset of units (the sample) from the whole population of interest so that a study can be run on the few while conclusions are made about the many. The examiner wants three things from a one-line vignette: name the sampling method, say whether it is **probability** or **non-probability**, and state whether the resulting sample can support a generalisation to the population (Bland 2015; Altman 1991; Kaplan & Sadock 2024). The single hinge is that only probability sampling, where every unit has a known non-zero chance of selection, permits valid statistical inference from sample to population.

The family splits cleanly. **Probability sampling** methods (simple random, systematic sampling, stratified, cluster sampling, multistage) use a chance mechanism and allow generalisable inference; **non-probability sampling** methods (convenience, purposive, snowball, quota) select by accessibility or judgement and cannot support a population estimate. Hold the two-column split and the rest is detail.

### 5.1 The population, the sampling frame, and the sample

**Three nested objects define every sampling problem.** The **target population** is the whole group about whom conclusions are wanted (for example, all adults in a district). The sampling frame is the actual list or operational device from which units are drawn (an electoral roll, a household census list, a clinic register): it is the population as it is really available to be sampled. The **sample** is the subset finally selected and measured (Bland 2015). Most real-world error enters at the frame: if the frame omits part of the target population (the homeless are missing from an electoral roll, the untreated are missing from a clinic register) the sample is biased before

a single unit is drawn, however elegant the selection method. Naming the frame is the first discipline of good sampling.

**Target population**
The whole group about whom inference is wanted.

**Sampling frame**
The actual list or device from which the sample is drawn; the practical stand-in for the population.

**Sampling unit**
The element selected at each draw (a person, a household, a school, a village).

### 5.2 Probability sampling: the licence to generalise

**Every unit has a known, non-zero chance of selection.** In **probability sampling** selection is by a chance mechanism, so the probability that any given unit enters the sample is known and greater than zero (Kaplan & Sadock 2024; Bland 2015). This is what makes a sample *representative in expectation* and lets sampling theory attach a standard error and a confidence interval to an estimate: only under a known selection probability can the mathematics of inference from sample to population be justified. Community psychiatric morbidity surveys therefore use probability sampling (the National Comorbidity Survey and the WHO World Mental Health surveys used multistage household probability sampling precisely to license generalisation to their target populations). If you cannot state each unit's chance of selection, you are not doing probability sampling and you forfeit the right to a population estimate.

RULE **Known chance, valid inference.** Only when every unit has a known non-zero probability of selection can a sample estimate be generalised to the population with a valid standard error and confidence interval.

### 5.3 Simple random and systematic sampling

**The two basic single-stage probability methods.** In **simple random** sampling every unit in the frame has an equal and independent chance of selection, drawn by lottery or random-number generator; it is the reference standard against which all other methods are judged and needs a complete numbered frame (Bland 2015; Altman 1991). **Systematic sampling** selects every kth unit from the ordered frame after a random start, where the sampling interval k equals the population size divided by the desired sample size (draw a random start between 1 and k, then take that unit, then every kth thereafter). Systematic sampling is quicker and needs no full random draw, and it spreads the sample evenly through the list; its one hazard is a hidden periodicity in the frame that coincides with the interval k (for example, if every kth bed is a corner bed), which can bias the sample. Both are probability methods and both give generalisable estimates.

**Simple random**
Equal, independent chance for every unit; the gold-standard probability method.

**Systematic sampling**
Every kth unit after a random start, k = population size / sample size; fast but vulnerable to periodicity in the frame.

**WORKED EXAMPLE**

You have a numbered frame of 1,000 clinic attenders and want a sample of 100. The sampling interval is k = 1000 / 100 = **10**. Pick a random start between 1 and 10, say 7, then take attender 7, 17, 27, 37 and so on to 997: that is **systematic sampling**. Every attender had a 1 in 10 (0.1) chance of selection, so it is a probability sample and the estimate generalises, provided the list carries no repeating pattern that lands on the interval of 10.

### 5.4 Stratified sampling: guaranteeing the subgroup

**Divide first into strata, then sample within each. Stratified** sampling first partitions the population into mutually exclusive strata defined by a characteristic that matters (age band, sex, urban versus rural, diagnosis), then draws a random sample *within each stratum* (Bland 2015; Altman 1991). Its whole purpose is to guarantee representation: because each stratum is sampled deliberately, even a small subgroup that a simple random draw might miss by chance is certain to appear, and stratifying on a variable related to the outcome also reduces the standard error of the overall estimate (it removes between-stratum variability). In proportionate stratification each stratum contributes in proportion to its share of the population; in disproportionate stratification a small but important stratum is deliberately over-sampled and the estimates are re-weighted at analysis (this is the oversampling of older adults or minority groups seen in the large epidemiological surveys). Stratified sampling never increases the standard error relative to simple random sampling and usually lowers it.

RULE **Stratify to represent a small subgroup.** When you must be certain a small but important subgroup appears in adequate numbers, stratify on that characteristic and sample within the stratum, oversampling it if needed and re-weighting at analysis.

### 5.5 Cluster sampling and the design effect

**Sample whole groups, not individuals. Cluster sampling** divides the population into naturally occurring groups (clusters) such as villages, schools, wards or households, takes a random sample of whole clusters, and then studies every unit within the selected clusters (or a further sample of them). It is chosen for logistics and cost: when the population is geographically dispersed and no complete individual-level frame exists, it is far cheaper to enumerate and travel to a handful of clusters than to chase a random scatter of individuals across the whole area (Bland 2015). The price is statistical. People within a cluster tend to resemble one another (they share environment, culture, exposures), so each additional person in a cluster adds less new information than an independent draw would. This within-cluster correlation inflates the variance, quantified by the design effect (Deff): the ratio of the variance under the cluster design to the variance a simple random sample of the same size would have given. A design effect greater than 1 means the cluster sample carries less information than its nominal size, so the sample size must be multiplied by the design effect to recover the intended precision. Stratified sampling reduces variance; cluster sampling increases it. That contrast is examined constantly.

#### Cluster sampling

Randomly sample whole clusters (villages, schools), then study units within them; cheaper for dispersed populations, no individual frame needed.

#### Design effect (Deff)

Variance under the cluster design divided by the variance a simple random sample would give; Deff > 1 means precision is lost and the sample must be enlarged.

TRAP **Ignoring the design effect.** Analysing a cluster sample as if it were a simple random sample understates the true standard error and gives falsely narrow confidence intervals and falsely small p-values; the sample size must be inflated by the design effect.

### 5.6 Multistage sampling: the real-world survey design

**Sampling in successive nested stages.** **Multistage** sampling combines the above into a sequence: sample large primary units first, then sub-sample within them, over two or more stages. A national survey might first randomly select districts (stage 1), then villages within selected districts (stage 2), then households within selected villages (stage 3), then one adult within each selected household (stage 4). It is the practical backbone of almost every large community survey because it needs a frame only for the units selected at the next stage down, never a complete individual-level frame for the whole country. Each clustering stage carries its own design effect, so multistage designs are analysed with survey-weighted methods that account for the stratification, the clustering and the unequal selection probabilities. The National Mental Health Survey type of design and the WHO World Mental Health surveys are multistage household probability samples of exactly this kind (Kaplan & Sadock 2024).

### 5.7 Non-probability sampling: convenience, purposive, snowball, quota

**Selection by access or judgement, not chance.** In **non-probability sampling** the chance of any unit being selected is unknown, so no valid generalisation to a population can be made; these methods belong to qualitative research, pilot studies and situations where a frame is impossible (Kaplan & Sadock 2024). **Convenience** sampling takes the units easiest to reach until the target size is met (a volunteer sample, the patients who happen to be in clinic today); it is fast and cheap but its self-selection makes it the least representative method. **Purposive** (judgemental) sampling selects units the researcher judges most informative for the question, the standard approach in qualitative work where the aim is depth and typicality, not statistical representativeness. **Snowball** sampling starts from a few index participants who then recruit further participants from their own networks, and is the method of choice for hidden or hard-to-reach populations (people who inject drugs, sex workers, the homeless) where no frame exists (Kaplan & Sadock 2024). Quota sampling fills preset quotas for chosen characteristics (so many men, so many women) by non-random selection until each quota is met: it superficially resembles stratified sampling but the crucial difference is that within each quota the units are chosen non-randomly, so it remains a non-probability method. Consecutive sampling, a near neighbour, recruits every accessible unit over a defined period.

TRAP **Convenience sampling cannot support a prevalence estimate.** A prevalence figure is a population statement; a convenience sample has an unknown selection probability, so it can describe only the sample and can never be quoted as the prevalence in the population.

### 5.8 Sampling error versus sampling bias

**Two different failures, only one of which shrinks with size.** Sampling error is the random difference between a sample estimate and the true population value that arises simply because a sample, not the whole population, was measured; it is unavoidable, it has no direction, and it is exactly what the standard error and the confidence interval quantify. Crucially, sampling error *decreases as the sample size increases* and disappears if the whole population is measured (Bland 2015). Sampling bias (selection bias) is a systematic error in the way units are selected, so that the sample is unrepresentative in a consistent direction (the classic example is drawing a sample of the mentally ill only from treated populations, which over-represents the more severe and help-seeking). Bias has a direction, it distorts the estimate the same way every time, and, fatally, *a larger sample does not fix it*: it merely produces a more precise wrong answer. Distinguishing the two, and knowing that size cures error but not bias, is one of the most reliably examined points in the whole topic.

| Random | Systematic |
| --- | --- |
| SAMPLING ERROR: NO DIRECTION, SHRINKS WITH N | SAMPLING BIAS: HAS DIRECTION, NOT FIXED BY N |

TRAP **A bigger sample does not cure bias.** Increasing the sample size narrows sampling error but a biased sampling method just gives a more precise wrong estimate; only a better sampling design removes bias.

### 5.9 The selector: matching the situation to the method

**Read the constraint, pick the method.** A sampling vignette turns on three things: is a complete frame available, must a particular subgroup be guaranteed, and is generalisation required at all? A complete individual frame invites simple random or systematic sampling; the need to represent a small subgroup calls for stratification; a dispersed population with no individual frame forces cluster or multistage sampling and its design effect; and a hidden population or a purely exploratory aim drops you into the non-probability methods, which forfeit generalisation.

| METHOD | PROBABILITY? | HOW IT SELECTS | WHEN TO USE IT |
| --- | --- | --- | --- |
| **Simple random** | Yes | Equal, independent chance for every unit (lottery / random numbers) | Complete numbered frame available; the reference standard |
| **Systematic sampling** | Yes | Every kth unit after a random start, k = N / n | Ordered frame; quicker than a full random draw; no periodicity in the list |

| METHOD | PROBABILITY? | HOW IT SELECTS | WHEN TO USE IT |
|---|---|---|---|
| **Stratified** | Yes | Split into strata, random-sample within each | Must guarantee a small subgroup; improves precision |
| **Cluster sampling** | Yes | Randomly sample whole clusters, study units within | Dispersed population, no individual frame; cheaper, but larger design effect |
| **Multistage** | Yes | Successive nested sampling stages | Large national surveys; needs a frame only stage by stage |
| **Convenience** | No | Whoever is easiest to reach | Quick pilots only; cannot generalise |
| **Purposive** | No | Researcher's judgement of who is informative | Qualitative work seeking depth and typicality |
| **Snowball** | No | Index participants recruit further participants | Hidden or hard-to-reach populations with no frame |
| **Quota** | No | Fill preset quotas by non-random selection | Fast field surveys; looks stratified but is not random |

The sampling family read across method, whether it is probability-based, how it selects, and when to use it; the hit column is the probability-versus-non-probability discriminator that decides generalisability.

**HOLD THIS**

Probability sampling (known non-zero selection chance) is the only route to a generalisable population estimate; stratify to guarantee a small subgroup, cluster to save cost at the price of a design effect greater than 1, and never quote a prevalence from a convenience sample.

**INDIA**

The National Mental Health Survey of India (2015 to 2016) is the flagship probability-sampled example: a **multistage** stratified random household survey across 12 states that could therefore report a generalisable current weighted prevalence of any mental disorder of about **10.6** percent and a treatment gap for common mental disorders exceeding **80** percent. Because selection probabilities were known and the design was multistage with clustering, the estimates carry valid confidence intervals: a convenience sample of clinic attenders could never have produced a national prevalence figure.

PEARL

Stratification and clustering pull the standard error in opposite directions. Stratified sampling *reduces* variance because it removes between-stratum variability, so its design effect is typically below 1; cluster sampling *increases* variance because units within a cluster are correlated, so its design effect is above 1. A survey that stratifies and then clusters, as most national surveys do, trades the two off, which is why the net design effect must be computed and the sample size inflated accordingly.

MNEMONIC

**Probability sampling: "Some Systems Sort Clusters Multiply."**

Simple random, Systematic, Stratified, Cluster, Multistage: the five probability methods. Everything else (convenience, purposive, snowball, quota) is non-probability and cannot generalise.

GOOD TO REMEMBER

- **Simple random:** every unit has an equal, independent chance of selection; use it whenever a complete numbered frame exists; it is the gold-standard probability method against which the design effect of others is judged (Deff = 1).
- **Systematic sampling:** take every kth unit after a random start; use it for a quick draw from an ordered list; the one number is the interval k = population size / sample size.
- **Stratified:** split into strata and random-sample within each; use it to guarantee a small subgroup is represented and to improve precision; its design effect is usually below 1 (it lowers the standard error).
- **Cluster sampling:** randomly sample whole clusters and study units within them; use it for a dispersed population with no individual frame to save cost; the number to know is the design effect (Deff greater than 1), by which the sample size must be multiplied.
- **Multistage:** sample in successive nested stages (districts, then villages, then households, then individuals); use it for large national surveys; it needs a frame only for the units at the next stage, and each clustering stage adds its own design effect.
- **Convenience:** take whoever is easiest to reach; use it only for quick pilots; the fact to hold is that its unknown selection probability means it cannot support a prevalence estimate.
- **Purposive:** select by the researcher's judgement of who is most informative; use it in qualitative research for depth and typicality; it makes no claim to statistical generalisation.
- **Snowball:** index participants recruit further participants through their networks; use it for hidden or hard-to-reach populations with no sampling frame; it is non-probability, so no population estimate follows.

## 6 · Qualitative research methods

**Some questions cannot be answered by counting.** **Qualitative** research is the systematic study of meaning and experience: it asks how people understand their illness, why an intervention fails in practice, what a service feels like from the inside. It collects and analyses data that cannot be reduced to numbers (interview transcripts, field notes, documents) and reads them for categories and themes rather than for means and p-values (Bowling 2014; Greenhalgh 2014). The exam wants three things from you: the named approaches with a one-line signature each, the concepts

that give a qualitative study its **rigour** (the qualitative counterpart of reliability and validity), and a clear sense of when a qualitative design beats a quantitative one. Get those and any stem, whether it names grounded theory or asks "what is saturation", falls open.

The mental model is a mirror image of the quantitative one. Where a trial samples randomly to represent a population and tests a pre-set hypothesis, a qualitative study samples **purposively** to capture informative cases and lets categories emerge from the data, with analysis running iteratively alongside collection (Bowling 2014). It generalises to theory and concepts, never to population statistics. Hold that contrast and the rest is detail.

### 6.1 What qualitative research is, and how it differs from quantitative

**Understanding meaning, not measuring frequency.** A **qualitative** study seeks to understand a phenomenon through personal encounter, observation and text, rather than to quantify it (Bowling 2014; Greenhalgh 2014). The defining differences from quantitative work are consistent and examinable: sampling is purposive (subjects chosen deliberately for characteristics relevant to the question, not random selection); data collection is open-ended and often interactive (interviews, observation, documents, not standardised scales); analysis is **iterative** and interpretive, generating categories and themes as the data accumulate rather than testing a fixed hypothesis at the end; and the aim is depth of understanding and **transferability** of concepts, not statistical generalisation to a population. The two are complementary, not rivals: a mixed-methods study can nest a qualitative arm inside a randomised trial to explain why participants responded as they did.

RULE **Purpose sets the design.** If the question is "how many / how much / does it work", think quantitative; if it is "why / how / what is it like", think qualitative.

### 6.2 The five main approaches

**Each approach has a distinct object and analytic move.** The examiner most often wants the named approach matched to its one-line signature (Bowling 2014; Greenhalgh 2014). Grounded theory builds a theory upward from the data themselves through **constant comparison**, comparing each new piece of data with the emerging categories until a theory grounded in the material is generated (Glaser and Strauss 1967). Thematic analysis is the workhorse method that codes the data and organises those codes into themes, without committing to a deeper theoretical framework (Braun and Clarke 2006). The phenomenological approach studies the **lived experience** of a phenomenon, the essence of what it is like to undergo it, from the first-person perspective. Ethnography studies a culture or group in its natural setting, classically through prolonged participant observation and immersion. Content analysis systematically categorises the content of text or documents, and can be handled qualitatively (interpreting meaning) or quantitatively (counting the frequency of coded categories).

#### Grounded theory

Theory built from the data by constant comparison; sampling continues until the theory is saturated.

#### Thematic analysis

Codes distilled into themes; flexible, framework-light, the commonest entry-level method.

**Phenomenological**
The lived experience and essence of a phenomenon, from the person's own perspective.

**Ethnography**
A culture or group studied in its natural setting through immersion and participant observation.

**Content analysis**
Systematic categorisation of textual or documentary content; qualitative or quantitative in flavour.

### 6.3 Data collection: focus groups and in-depth interviews

**Two workhorses, chosen by what you want to hear.** The two staple methods of qualitative data collection are the **focus group** and the in-depth interview (Bowling 2014). A focus group brings a small group of participants (typically six to ten) together with a moderator so that the *interaction between members* generates data: people react to and build on each other's views, which surfaces shared meanings, norms and disagreements that a one-to-one setting would miss. Its weakness is that dominant voices can silence others and it is poor for private or sensitive material. In-depth (semi-structured or unstructured) interviews put one participant with one interviewer using an open topic guide, giving privacy and depth on an individual's experience, and are the natural choice for sensitive subjects or for reconstructing a personal account. Both are complemented by participant observation (watching behaviour in its natural context) and by documentary sources. Sampling for all of these is purposive, and often theoretical (sampling driven by the emerging analysis) in grounded theory.

### 6.4 Saturation and when to stop sampling

**Stop when new data stop teaching you anything.** Because a qualitative sample is chosen for information richness rather than for a pre-calculated power, its size is not fixed in advance; it is governed by saturation. **Saturation** is the point at which no new codes, categories or themes emerge from further data collection, so additional interviews or observations add nothing to the analysis (Bowling 2014). Reaching it is the qualitative analogue of an adequate sample size: you keep sampling and analysing in parallel and stop once the emerging framework has stabilised. This is why qualitative sample sizes are small and cannot be stated at the outset the way a trial's n can.

RULE **Stop sampling at saturation.** Continue collecting and analysing data until no new codes or themes emerge; that point, not a pre-set number, sets the qualitative sample size.

### 6.5 Rigour: the trustworthiness criteria and the techniques that build them

**The qualitative counterpart of reliability and validity.** A good qualitative study is judged not by statistical validity but by its **trustworthiness**, the framework of Lincoln and Guba (1985) with four criteria: credibility (are the findings a believable account of the participants' reality, the qualitative parallel of internal validity), transferability (can the findings be applied to other settings, the parallel of external validity, supported by rich "thick description" so a reader can judge fit), dependability (would the process be consistent and traceable if repeated, the parallel of reliability), and confirmability (are the findings grounded in the data rather than the researcher's

bias, the parallel of objectivity). Several techniques operationalise these criteria (Bowling 2014; Greenhalgh 2014):

**Triangulation**

Triangulation: using multiple sources, methods, investigators or theories and checking that they converge on the same finding; convergence strengthens credibility.

**Reflexivity**

Reflexivity: the practice of reflexivity means the researcher explicitly examines and discloses how their own values, assumptions and presence shape the data and its interpretation.

**Member checking**

Member checking (respondent validation): taking the findings back to participants to confirm they recognise their experience in the account; supports credibility.

**Audit trail**

An audit trail: a transparent, documented record of every decision from raw data to conclusion, so an outsider can follow the reasoning; supports dependability and confirmability.

TRAP **Qualitative findings are not statistically generalisable.** They transfer to theory and to similar settings (transferability), but they cannot be projected onto a population the way a representative quantitative estimate can.

### 6.6 When qualitative beats quantitative

**Reach for qualitative when the number is not the answer.** Qualitative methods are the right tool, not a weaker substitute, in defined situations (Bowling 2014; Greenhalgh 2014). Use them for **hypothesis generation**, where little is known and the aim is to discover the relevant concepts and variables before any quantitative study can be designed. Use them to understand **barriers** and facilitators, why an evidence-based intervention is not adopted in practice, why patients disengage, what obstructs help-seeking, since these turn on meaning and context that a scale cannot capture. Use them for **service evaluation**, to gather the attitudes, beliefs and experiences of service users, carers and professionals, and to see how a service actually works on the ground. They also excel at exploring sensitive or stigmatised experiences and at explaining the "why" behind a surprising quantitative result.

### 6.7 The selector: matching approach and method to the question

**Read the aim, then pick the approach.** Faced with a stem, first confirm the question is qualitative (a why/how/what-is-it-like question), then match the specific aim to the approach and the setting to the collection method. If the aim is to build an explanatory theory, choose grounded theory; if it is simply to identify patterns of meaning, thematic analysis; if it is the essence of an experience, phenomenology; if it is a group's culture in situ, ethnography; if it is the systematic content of documents, content analysis. The collection method then follows from whether you need group interaction or private depth.

| AIM OF THE STUDY | APPROACH / METHOD TO PICK |
|---|---|
| Build a new explanatory theory from the ground up | Grounded theory (constant comparison) |
| Identify patterns of meaning across the data | Thematic analysis (codes to themes) |
| Capture the essence of a lived experience | Phenomenological approach |
| Understand a group's culture in its natural setting | Ethnography (participant observation) |
| Systematically categorise documents or text | Content analysis |
| Surface shared norms and group interaction | Focus group |
| Explore a private or sensitive individual account | In-depth interview |

Qualitative selector: read the study aim, pick the approach or data-collection method (aim to method).

**WORKED EXAMPLE**

A team wants to know why patients started on a depot antipsychotic stop attending the clinic. A trial can measure the drop-out rate, but not the reasons. They run in-depth interviews with disengaged patients and two focus groups with community nurses, code the transcripts, and organise the codes into themes (a thematic analysis). They keep interviewing until no new theme appears (saturation), take the draft themes back to a few patients to check they ring true (member checking), and confirm that the interview, the nurse focus groups and the case notes all point to the same barriers (triangulation). The output is not a percentage but a set of transferable themes ("feeling talked down to", "side effects not discussed") that a service can act on, and that can seed a later quantitative study.

**HOLD THIS**

Qualitative research studies meaning and experience, samples purposively, and stops at saturation; its four trustworthiness criteria are credibility, transferability, dependability and confirmability, and its findings transfer to theory, not to population statistics.

**COMMON TRAP**

Two errors examiners set. First, treating qualitative findings as if they were statistically generalisable to a population: they are not, they offer transferability to similar contexts, judged from thick description, not statistical generalisation. Second, criticising a qualitative study for a "small, unrepresentative sample": small purposive samples are correct by design, and adequacy is judged by saturation, not by size or representativeness. Judge a qualitative study by its trustworthiness, not by the yardstick of a trial.

MNEMONIC

**Trustworthiness = "Can These Data Convince" (Credibility, Transferability, Dependability, Confirmability).**

The four Lincoln and Guba criteria map onto the quantitative quartet: Credibility to internal validity, Transferability to external validity, Dependability to reliability, Confirmability to objectivity. "Can These Data Convince" fixes the order and reminds you the whole point is a believable, traceable account.

GOOD TO REMEMBER

- **Grounded theory:** builds theory upward from data; use it to generate a new explanatory model; the one move to know is constant comparison, sampling to theoretical saturation.
- **Thematic analysis:** the flexible framework-light method; use it to find patterns of meaning; the one move is coding the data then grouping codes into themes.
- **Phenomenological:** studies the lived experience and essence of a phenomenon; use it to capture what an experience is like from the inside.
- **Ethnography:** studies a culture or group in its natural setting; use it for immersion and participant observation over time.
- **Content analysis:** systematically categorises text or documents; use it on documentary data; it can be qualitative (meaning) or quantitative (frequency counts).
- **Focus group:** six to ten participants with a moderator; use it when the interaction between members is the data; weak for private or sensitive material.
- **Saturation:** the stopping rule; the point at which no new codes or themes emerge; it, not a pre-set n, sets qualitative sample size.
- **Triangulation:** convergence of multiple sources, methods or investigators on one finding; the number to state is at least two independent sources agreeing.
- **Trustworthiness (Lincoln and Guba):** credibility, transferability, dependability, confirmability; the qualitative parallels of internal validity, external validity, reliability and objectivity.

## 7 · Systematic review and meta-analysis

**This is the apex of the evidence pyramid.** A single trial answers one question in one sample; the **systematic review** gathers every trial that ever asked that question and reads them together, and the **meta-analysis** is the statistics that pool their numbers into one estimate. The examiner treats these as the most senior methods in the paper, and the marks sit in two places: the distinction between the reproducible **systematic review** method and the older narrative review, and the interpretation of the forest plot with its diamond, its heterogeneity and its funnel-plot check for publication bias (Shorter Oxford Textbook of Psychiatry; Companion to Psychiatric Studies 2010).

Read the topic as one pipeline. A protocol fixes the question and the search; a comprehensive search finds the trials; explicit criteria include or exclude them; risk-of-bias appraisal weighs each; and only then, optionally, a meta-analysis pools the results, tests them for heterogeneity, checks the literature for publication bias, and grades the certainty of the whole. The systematic review is the method; the meta-analysis is the arithmetic that may or may not sit inside it (Bland 2015).

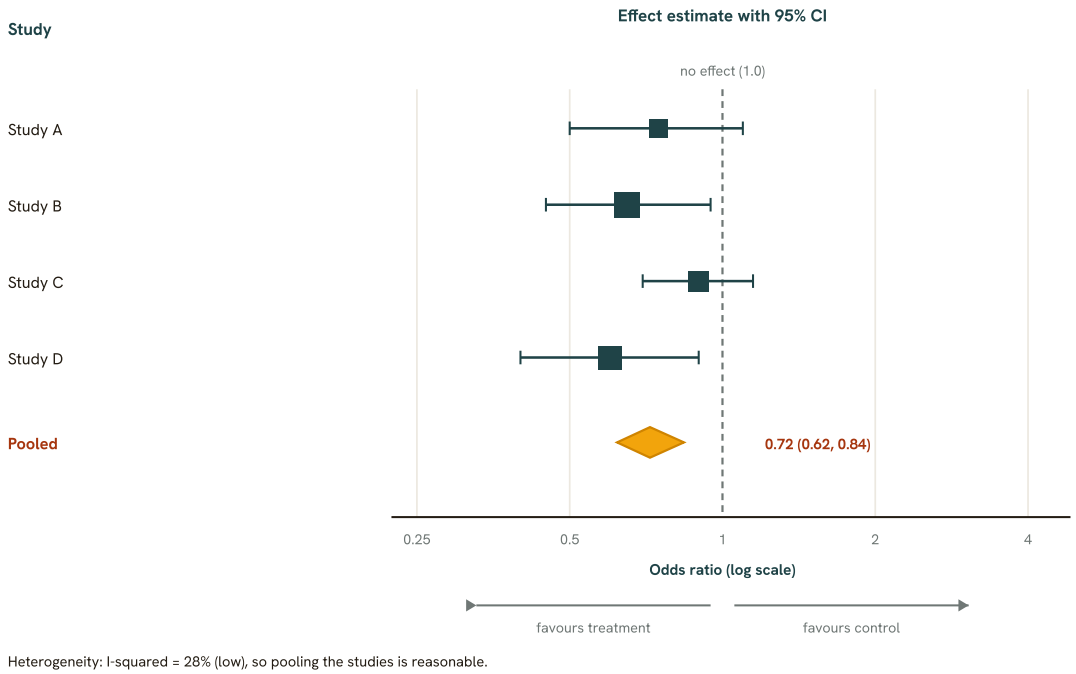



**Fig 7.3** Reading a forest plot. Each study is a horizontal line (its 95% confidence interval) with a box marking the point estimate, the box sized by the study's weight in the pool. The vertical dashed line is the value of no effect. A confidence interval that crosses it is non-significant for that study. The diamond at the foot is the pooled estimate, its width the pooled confidence interval. Before trusting the pool, check heterogeneity: a high I-squared means the studies disagree and a single summary may mislead.

## 7.1 The systematic review: a protocol-driven synthesis, not a narrative review

**A systematic review is a reproducible, protocol-driven summary of all the evidence on one question.** A systematic review starts from a clearly formulated question, then follows a pre-registered protocol: a comprehensive and documented search, explicit and pre-stated inclusion and exclusion criteria, formal risk-of-bias (quality) appraisal of each included study, and a transparent synthesis. Because every step is specified in advance and reported, a second team following the same protocol should reach the same set of studies and the same conclusion, which is precisely what makes it reproducible (Shorter Oxford Textbook of Psychiatry). This is the opposite of the traditional **narrative review**, in which an expert selects and weights the literature by personal judgement, so the reader cannot tell what was searched, what was left out, or why. The two-investigator rule, that study selection and quality scoring be done independently by at least two reviewers, exists to keep even the systematic review honest.

**RULE** **Systematic review = method; meta-analysis = optional statistics.** A systematic review is defined by its protocol (question, search, criteria, appraisal); the pooling is a separate step that it may or may not include.

**TRAP** **Not every systematic review contains a meta-analysis, and not every pooled analysis is systematic.** If the trials are too clinically diverse to combine, a good systematic review synthesises them narratively and refuses to pool; pooling them anyway is the error.

### 7.2 The Cochrane Collaboration and the evidence hierarchy

**The systematic review of randomised trials sits at the top of the therapy hierarchy.** The movement grew from Archibald Cochrane, whose 1972 book argued that treatment provision should rest on randomised controlled trials because they are more reliable than any other evidence (Cochrane 1972). His call for a critical, periodically updated summary of all relevant trials was realised by the Cochrane Collaboration, formed in 1993 and now the largest organisation in the world producing and maintaining systematic reviews (Shorter Oxford Textbook of Psychiatry). In the conventional hierarchy of evidence for questions of therapy, level Ia is a systematic review of randomised controlled trials and level Ib is at least one randomised controlled trial, both above non-randomised controlled studies (II), descriptive studies (III), and expert opinion (IV). The **cochrane** review, with its rigorous protocol and standard software, is the archetype of the method.

**INDIA**

Cochrane reviews are freely accessible in India through the Cochrane Library and are the standard evidence base cited by Indian clinical practice guidelines. When guidance from a Western meta-analysis is read across to Indian practice, the external-validity question of whether the pooled sample resembles Indian patients still applies, because most pooled trials are from high-income settings.

### 7.3 Meta-analysis: the statistical pooling, distinguished from the systematic review

**Meta-analysis is the quantitative pooling of results across studies into one weighted estimate.** A meta-analysis takes the effect estimate from each included study (an odds ratio or relative risk for a binary outcome, a mean difference or standardised effect size for a continuous one) and combines them into a single pooled estimate of greater precision than any one trial (Shorter Oxford Textbook of Psychiatry). Crucially, it does not weight studies equally: each contributes in proportion to its information, so larger and more precise studies pull the pooled estimate more (the inverse-variance principle). A meta-analysis is therefore a step that lives inside a systematic review, never a substitute for it: pooling numbers gathered without a systematic, unbiased search simply averages a biased sample of the literature. The systematic review supplies the trustworthy set of studies; the meta-analysis does the arithmetic on them.

**WORKED EXAMPLE**

Three trials of a drug versus placebo report relative risks of relapse of 0.80, 0.70 and 0.75, each with its own confidence interval. A meta-analysis does not average these as (0.80 + 0.70 + 0.75) / 3; it weights each by the inverse of its variance, so the largest, most precise trial dominates, yielding a pooled relative risk near **0.75**, its exact value set by the study weights, with a confidence interval narrower than any single trial. The narrowing of the confidence interval, not a new point estimate, is the gain from pooling.

### 7.4 Reading the forest plot

**The forest plot displays every study, its weight, and the pooled result on one axis.** The forest plot is the signature graphic of a meta-analysis, and reading it is a standard viva task (Shorter

Oxford Textbook of Psychiatry). Each horizontal line is one study: the central box is its point estimate, and the size of the box marks its weight in the pool (larger box = more precise study, more influence), while the horizontal whiskers are its 95% confidence interval. A vertical **null line** runs down the plot at the value of no effect: **1.0** for a ratio measure (odds ratio, relative risk) and **0** for a difference measure (mean difference). A study whose confidence interval crosses the null line is individually non-significant. At the foot of the plot sits the pooled result as a diamond: its centre is the pooled point estimate and its width is the pooled confidence interval, so if the diamond touches or crosses the null line the overall result is non-significant.

**Box (point estimate)**
Each study's own effect; the box centre is the estimate, the box size is that study's weight in the pool.

**Whiskers**
The 95% confidence interval of that single study; if they cross the null line, that study alone is not significant.

**Null line**
The vertical line of no effect: 1.0 for odds ratio or relative risk, 0 for a mean difference.

**Diamond**
The pooled estimate at the bottom: centre is the combined effect, width is its confidence interval; crossing the null line means the pooled result is non-significant.

### 7.5 Fixed-effect versus random-effects models

**The choice of model depends on whether the studies share one true effect.** The pooling can assume one of two things. A fixed-effect model assumes every study is estimating the same single underlying treatment effect, and all variation between them is sampling error; it weights purely by study size and gives the narrowest, most confident pooled confidence interval. A random-effects model assumes the true effect genuinely differs from study to study (because of differing doses, patients or settings), so it adds this between-study variance to the weighting; it gives a wider, more conservative confidence interval and lets small studies count for relatively more (Shorter Oxford Textbook of Psychiatry). The decision follows the heterogeneity: when studies are consistent, a fixed-effect model is appropriate and efficient; when significant heterogeneity is present, the random-effects model is the more appropriate and honest choice, and its wider interval reflects the real uncertainty.

RULE **Significant heterogeneity means use a random-effects model.** Fixed-effect assumes one shared true effect; when effects genuinely vary, random-effects widens the interval to tell the truth about that variation.

### 7.6 Heterogeneity: the chi-square test and the I-squared statistic

**Heterogeneity is real variation in the true effect between studies, beyond chance.** When some trials show a large effect and others none at all, the studies show heterogeneity, meaning their results are more different than sampling error alone can explain (Shorter Oxford Textbook of Psychiatry). It is assessed two ways. The **chi-square test for heterogeneity** (Cochran's Q, a modification of the chi-square test) formally tests the null hypothesis that all studies share one

true effect; because it has low power when trials are few, a lenient threshold such as p less than **0.10** is conventionally used, and a significant result signals heterogeneity. The **I-squared** statistic (written **i-squared** or **i2**) is the more useful companion: it is the percentage of total variation across studies that is due to real heterogeneity rather than chance, running from 0% to 100% and, unlike the chi-square, not dependent on the number of studies. The rough Cochrane bands are worth holding.

| 0 to 40% | 30 to 60% | 50 to 90% | 75 to 100% |
|---|---|---|---|
| I2: MIGHT NOT BE IMPORTANT | I2: MODERATE | I2: SUBSTANTIAL | I2: CONSIDERABLE |

Overlapping I-squared bands from the Cochrane Handbook (Higgins et al 2023); the overlap is deliberate, since the importance of a given value also depends on the direction and strength of the evidence.

TRAP **A tidy pooled diamond over high heterogeneity is misleading.** High i-squared means the studies are answering different questions; the neat pooled estimate then averages apples and oranges and should not be reported as if it were one clean effect.

### 7.7 Publication bias: the funnel plot and Egger's test

**Positive studies are published more, so the pooled literature can overstate the effect.** Publication bias is the tendency for statistically significant, positive trials to be submitted and published more readily than null or negative ones, so a meta-analysis restricted to published studies systematically exaggerates the benefit (Shorter Oxford Textbook of Psychiatry). The classic detection tool is the funnel plot: each study's effect estimate is plotted against its precision (or standard error), so small imprecise studies scatter widely at the bottom and large precise ones cluster tightly at the top. With no bias the cloud is a symmetric inverted funnel about the pooled estimate; **funnel plot** asymmetry, a corner of small negative studies missing, suggests that such studies were run but never published. Because eyeballing asymmetry is subjective, Egger's test provides a formal statistical test of funnel-plot asymmetry (Egger et al 1997). The remedies are structural: trial registries, and searching grey literature and unpublished data.

COMMON TRAP

Funnel-plot asymmetry is not proof of publication bias. Genuine heterogeneity, poorer methodology in smaller trials, or chance can all bend the funnel. Asymmetry is a prompt to investigate, not a verdict; and a funnel plot needs roughly ten or more studies before its shape can be judged at all.

### 7.8 PRISMA reporting and GRADE certainty

**PRISMA standardises the reporting; GRADE rates how much to trust the result.** Two frameworks bracket the modern systematic review. The **PRISMA** statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standardises what must be reported about how the review was searched, conducted, analysed and presented, including the flow diagram of records screened, excluded and included (Moher et al 2009). Separately, GRADE (Grading of Recommendations Assessment, Development and Evaluation) rates the certainty of the pooled evidence for each outcome as high, moderate, low or very low. It starts a body of

randomised evidence at high certainty and rates it down for five problems, risk of bias, inconsistency (heterogeneity), indirectness, imprecision and publication bias, and can rate observational evidence up for a large effect or dose-response. GRADE is thus the summary judgement on top of the whole pipeline: a meta-analysis can produce a precise diamond and still be graded low certainty if the trials were biased or heterogeneous.

**PEARL**

Keep the three frameworks in their lanes. PRISMA governs reporting (how the review is written up), the risk-of-bias tool (such as the Cochrane RoB tool) appraises each included trial, and GRADE rates the certainty of the pooled evidence per outcome. A review can be PRISMA-compliant in its reporting yet GRADE-low in its certainty; good reporting of weak evidence is still weak evidence.

### 7.9 How the pipeline holds together

**Each stage protects the next, and the meta-analysis is only as good as the review beneath it.** The protocol fixes an unbiased question and search; the comprehensive search and explicit criteria assemble a fair set of studies; risk-of-bias appraisal weighs their quality; the meta-analysis pools them, but only after heterogeneity is checked so that pooling is legitimate; the funnel plot and Egger's test interrogate the literature for missing negative trials; and GRADE delivers the final verdict on certainty. Assess any published meta-analysis by walking this chain: an impressive pooled diamond built on a narrow search, unappraised trials, or ignored heterogeneity is exactly where the exam wants you to point.

**HOLD THIS**

The systematic review is the reproducible protocol-driven method (question, search, criteria, appraisal); the meta-analysis is the optional inverse-variance pooling inside it, read off a forest plot whose box is each study's weight, whisker is its CI, null line is 1.0 (or 0), and bottom diamond is the pooled estimate; high i-squared means use a random-effects model and distrust a tidy diamond, and a lopsided funnel plot (tested by Egger's) flags publication bias.

**GOOD TO REMEMBER**

- **Systematic review:** a reproducible synthesis of all evidence on one question; use it whenever a therapy question needs the whole literature, not one trial; defined by its protocol (question, comprehensive search, explicit inclusion, risk-of-bias appraisal), which separates it from a narrative review.
- **Cochrane Collaboration:** the world's largest producer of systematic reviews (formed 1993, from Archibald Cochrane's 1972 argument); the archetype of the method; its review of randomised trials is level Ia, the top of the therapy evidence hierarchy.
- **Meta-analysis:** the statistical pooling of study results into one weighted estimate; use it inside a systematic review when studies are similar enough to combine; weights each study by the inverse of its variance, so the gain is a narrower pooled confidence interval, not a new average.
- **Forest plot:** the display of a meta-analysis; read it to see each study and the pooled result; box = point estimate and weight, whisker = 95% CI, null line = 1.0 for a ratio (0 for a difference), bottom diamond = pooled estimate and its CI.
- **Fixed-effect vs random-effects model:** fixed assumes one shared true effect (narrower CI); random assumes effects vary between studies (wider, more conservative CI); use random-effects when heterogeneity is significant.
- **Heterogeneity (i-squared, chi-square):** real between-study variation beyond chance; the chi-square (Cochran's Q) test flags it at p less than 0.10, and i-squared gives the percentage of variation due to heterogeneity (bands: 0 to 40% low, 30 to 60% moderate, 50 to 90% substantial, 75 to 100% considerable).
- **Publication bias (funnel plot, Egger's test):** positive studies published preferentially, so the pool overstates effect; plot effect against precision, look for an asymmetric funnel (missing small negative studies), and test asymmetry formally with Egger's test; needs roughly ten or more studies.
- **GRADE:** rates the certainty of the pooled evidence per outcome as high, moderate, low or very low; starts randomised evidence high and rates down for risk of bias, inconsistency, indirectness, imprecision and publication bias; the final judgement on top of the whole review.

## 8 · Reporting guidelines and research ethics

Two examiner favourites sit back to back here: how a finished study should be *reported* so a reader can appraise it, and how a study is allowed to be *done* to human beings in the first place. The first half is a matching game, match the **reporting guideline** to the study design, and it underpins every structured critical appraisal. The second half is the ethical spine, from Nuremberg to the four principles to the consent process to the committee that signs off. Vivas love the crisp discriminators: which checklist for a trial versus a review, which document introduced "respect for persons", why consent is a process and not a signature.

**Frame it as two questions.** "Is this paper reported well enough to trust?" is answered by the **reporting guideline** matched to the design. "Was this study ethical to run?" is answered by the historical codes, the four principles, valid **informed consent**, and clearance by an **ethics committee**. Keep them separate and the topic collapses into two short tables.

### 8.1 Critical appraisal: reading a paper against a checklist

**What it is.** Critical appraisal is the structured process of judging whether a study's results are valid, what they mean, and whether they apply to your patients (Greenhalgh 2019). It weighs internal validity (was the study done well) against external validity (does it generalise). Reporting

guidelines exist so that authors give you every item you need to make that judgement, so a well-reported paper is not automatically a good study, but a badly reported one cannot be appraised at all. This is why a critical appraisal always begins by asking which reporting guideline the design demanded, then checking the paper against it.

RULE **Appraise before you apply.** Judge validity, size of effect, and applicability in that order; a statistically significant result from a biased study still does not change practice.

### 8.2 Reporting guidelines matched to design

**The matching game.** Each major design has its own EQUATOR-network checklist. The three you must never confuse: CONSORT for randomised trials, PRISMA for systematic reviews and meta-analyses, and STROBE for observational studies (Schulz et al. 2010; Moher et al. 2009; von Elm et al. 2007). Two more are named constantly alongside them: GRADE is not a reporting checklist but a system for rating the *certainty* of a body of evidence and the strength of a recommendation (Guyatt et al. 2008), and MOOSE is the reporting guideline for meta-analyses of *observational* studies where PRISMA (built for reviews of trials) is a poor fit (Stroup et al. 2000).

| GUIDELINE | DESIGN IT REPORTS | ONE-LINE DISCRIMINATOR |
|---|---|---|
| **consort (CONSORT)** | Randomised controlled trials | **Has the flow diagram: enrolled → randomised → analysed** |
| **prisma (PRISMA)** | Systematic reviews and meta-analyses (of trials) | **Has the four-phase flow: identified, screened, eligible, included** |
| **strobe (STROBE)** | Observational: cohort, case-control, cross-sectional | **Default for any study with no randomisation by the investigator** |
| **moose (MOOSE)** | Meta-analyses of observational studies | **Observational meta-analysis, where PRISMA does not fit** |
| **grade (GRADE)** | Certainty of a body of evidence + strength of recommendation | **Rates evidence high/moderate/low/very-low, not a reporting checklist** |

Match the guideline to the design; CONSORT, PRISMA, STROBE are the three that get swapped in vivas.

RULE **CONSORT for trials, PRISMA for reviews, STROBE for observational.** Name the design first, then the checklist follows automatically.

TRAP **PRISMA is for the review, not the trials in it.** A meta-analysis of RCTs is reported with PRISMA; each included RCT was reported with CONSORT. GRADE then rates the certainty of the pooled evidence. Do not answer "CONSORT" for a meta-analysis.

**GOOD TO REMEMBER**

- **CONSORT:** reporting guideline for randomised trials; use when appraising an RCT; the signature item is the participant flow diagram (enrolled, randomised, lost, analysed).
- **PRISMA:** reporting guideline for systematic reviews and meta-analyses; use for a pooled synthesis of studies; the signature item is the four-phase study-selection flow diagram.
- **STROBE:** reporting guideline for observational studies; use for cohort, case-control, or cross-sectional designs; the rule is "no investigator-controlled randomisation, use STROBE".
- **MOOSE:** reporting guideline for meta-analyses of observational studies; use when the pooled studies are observational, not trials; distinguishes itself from PRISMA by the observational input.
- **GRADE:** system to rate certainty of evidence and strength of recommendation; use when grading a body of evidence for a guideline; the four certainty levels are high, moderate, low, very low.

## 8.3 Research ethics: the historical spine

**Why the codes exist.** Research ethics is a reaction to abuse. Every modern safeguard traces to a documented harm, and examiners want the three landmark documents in order with what each one introduced (Emanuel et al. 2000; Kaplan & Sadock 2024).

### Nuremberg Code (1947)

Written after the Nazi doctors' trial; the first international code; its cardinal rule is that voluntary consent of the human subject is absolutely essential.

### Declaration of Helsinki (1964, WMA, revised repeatedly)

The physician-authored standard for medical research; introduced the research ethics committee, the written protocol, and the primacy of the participant's interests over those of science and society.

### Belmont Report (1979, USA)

Followed the Tuskegee syphilis study; named three foundational principles, respect for persons, beneficence, and justice, and mapped them to consent, risk-benefit assessment, and fair subject selection.

**MNEMONIC**

### Never Have Bad Research

Nuremberg → Helsinki → Belmont, the chronological spine. Nuremberg gave consent, Helsinki gave the ethics committee and protocol, Belmont gave the principles that became the four-principle framework.

## 8.4 The four principles

**The Beauchamp and Childress framework.** Biomedical ethics is standardly taught as four principles that must be balanced against each other (Beauchamp & Childress 2019; Kaplan & Sadock 2024). Belmont named three; the fourth, non-maleficence, is usually split out from beneficence in the clinical version.

### Autonomy

Respect for the person's right to make their own informed, voluntary decision; the ground of informed consent and confidentiality.

### Beneficence

Act in the participant's or patient's best interest; maximise benefit.

**Non-maleficence**

"First, do no harm"; minimise and justify risk; favourable risk-benefit ratio.

**Justice**

Fair distribution of the burdens and benefits of research; do not exploit vulnerable groups nor exclude those who could benefit.

PEARL

The Belmont-to-four-principles map is a classic one-mark question: respect for persons becomes autonomy, beneficence stays beneficence (and yields non-maleficence), and justice stays justice. Belmont's three "applications" (informed consent, risk-benefit assessment, selection of subjects) map onto the same principles.

### 8.5 Informed consent: capacity, disclosure, voluntariness

**The three elements.** Valid informed consent requires three things together (Kaplan & Sadock 2024; Appelbaum 2007). **Capacity** (the clinical term; "competence" is the legal, court-determined term) is the ability to communicate a choice, understand the relevant information, appreciate its application to oneself, and reason with it. **Disclosure** means giving the information a reasonable patient (patient-based standard) or reasonable physician (professional standard) would need: nature, purpose, risks, benefits, and alternatives including no treatment. **Voluntariness** means the decision is free of coercion, manipulation, or undue influence.

TRAP **Consent is a process, not a signature.** A signed form with no genuine understanding is not valid consent; consent is ongoing and can be withdrawn at any time. The form documents the process, it does not constitute it.

GOOD TO REMEMBER

- **Capacity:** the clinical ability to decide; four abilities, communicate a choice, understand, appreciate, reason; "capacity" is clinical, "competence" is the legal adjudication.
- **Disclosure:** the information given; must cover nature, risks, benefits, and alternatives including the option of no treatment; standard is reasonable-patient or reasonable-physician.
- **Voluntariness:** a free choice; invalidated by coercion, manipulation, or undue influence.

### 8.6 Vulnerable groups, surrogate consent, and assent

**When the participant cannot consent alone.** Some groups cannot give fully autonomous consent and need extra protection: children, those with impaired capacity (intellectual disability, dementia, acute psychosis), prisoners, and the economically or educationally disadvantaged where consent may not be truly voluntary. Two mechanisms cover them. **Surrogate consent** is proxy consent by a legally authorised representative (a legally acceptable representative or LAR) who decides in the person's best interest when the participant lacks capacity. **Assent** is the affirmative agreement of a minor or a person lacking full capacity who cannot legally consent; the child's assent is sought *in addition to* the parent's or guardian's consent, and a clear refusal (dissent) is generally respected in non-therapeutic research.

RULE **Consent from the guardian, assent from the minor.** A minor cannot legally consent; obtain the guardian's consent and the child's assent, and honour the child's dissent where the research offers no direct benefit.

GOOD TO REMEMBER

- **Surrogate (proxy) consent:** consent given by a legally authorised representative for a participant who lacks capacity; use for adults unable to decide; the standard is the participant's best interest.
- **Assent:** a minor's affirmative agreement alongside guardian consent; use for children and others lacking legal capacity; a documented refusal (dissent) is respected in non-beneficial research.

### 8.7 Governance: the ethics committee and institutional review board

**The gatekeeper.** No human research proceeds without prior clearance by an independent committee, called an ethics committee (EC, the common term in the UK, India, and Europe) or an institutional review board (IRB, the US term). Its job is to protect the rights, safety, and wellbeing of participants: it reviews the protocol, the risk-benefit balance, the consent process and documents, compensation, and provisions for vulnerable groups, and it continues to monitor the study through progress reports and adverse-event reporting. A conflicted or poorly reported study is stopped here before a single participant is enrolled.

GOOD TO REMEMBER

- **Ethics committee / institutional review board:** the independent body that reviews and monitors human research; use before enrolment and throughout; its remit is participant rights, safety, and wellbeing (protocol, risk-benefit, consent, vulnerable-group safeguards).

### 8.8 The Indian frame: ICMR National Ethical Guidelines 2017

**The national standard.** In India, biomedical and health research on human participants is governed by the ICMR National Ethical Guidelines for Biomedical and Health Research Involving Human Participants, 2017, which replaced the 2006 edition. They mandate prior review by a registered ethics committee, informed consent, additional safeguards for vulnerable participants, and are the reference standard for Indian institutional ethics committees. New Drugs and Clinical Trials Rules 2019 (under CDSCO) add the statutory regulatory layer for drug trials.

INDIA

Cite the ICMR National Ethical Guidelines 2017 for any Indian human-research question; they are the exam's expected answer for ethics-committee review and consent standards in India, superseding the 2006 guidelines. For a drug trial, add the New Drugs and Clinical Trials Rules 2019 (CDSCO). Registration of the ethics committee is mandatory.

### 8.9 Putting it together in an appraisal

**The examiner's synthesis.** A complete critical appraisal reads on two axes at once. On the reporting axis: name the design, pick the reporting guideline (CONSORT / PRISMA / STROBE), check the paper against it, and if it is a body of evidence for a recommendation, apply GRADE.

On the ethics axis: confirm ethics committee approval, valid informed consent, and appropriate handling of any vulnerable participants. A paper that is beautifully reported but ethically unsound, or ethically clean but unappraisable, fails.

**COMMON TRAP**

Confusing GRADE with a reporting checklist. GRADE rates the *certainty* of evidence (high, moderate, low, very low) and the strength of a recommendation; it is not the checklist you use to report a single study. Report a single RCT with CONSORT; grade the pooled evidence with GRADE.

**HOLD THIS**

CONSORT reports randomised trials, PRISMA reports systematic reviews and meta-analyses, STROBE reports observational studies; GRADE rates certainty of evidence, MOOSE reports observational meta-analyses; and informed consent is capacity plus disclosure plus voluntariness, a continuing process and never merely a signature.

## Biostatistics

### 9 · Data types and scales of measurement

**Before any statistic is computed, one question decides which analysis is legal: what kind of number is this?** The **scales of measurement** are the classification that governs which summary statistic and which statistical test may lawfully be applied to a variable. Every stem that hands you a variable ("tumour stage", "temperature in Celsius", "number of admissions", "diagnosis") is really asking you to name its scale, because the scale, not the researcher's preference, fixes whether you may quote a mean or only a median, and whether a parametric or a non-parametric test is permitted (Bland 2015; Altman 1991; Kaplan & Sadock 2024). Get the scale right and the correct description and test follow almost automatically.

The classification runs on two axes. First, Stevens' four **scales of measurement** in ascending order of information: **nominal**, **ordinal**, **interval**, **ratio** (Stevens 1946). Second, the broader **categorical** versus numerical split, with numerical data further divided into **discrete** and **continuous**. Nominal and ordinal are the two categorical scales; interval and ratio are the two numerical scales. Hold that mapping and the whole topic is scaffolded.

#### 9.1 Why the scale is the first decision, not an abstraction

**The scale determines the permitted arithmetic.** Distinguishing the **data types** may look like a needless abstraction, but it is essential in fixing the correct method of both data description and analysis (Companion to Psychiatric Studies; Bland 2015). What operations are meaningful climbs with the scale: on a nominal scale you may only count and compare for equality; on an ordinal scale you may additionally rank; on an interval scale you may add and subtract differences because the gaps are equal; on a ratio scale you may also multiply and form ratios because a true zero exists. A mean requires that differences be equal, so it is meaningful only from interval level

upward; below that, the median or the mode is the honest centre. This single dependency, arithmetic permitted rises with scale, is the engine behind every choice of statistic and test that follows.

RULE **The measurement scale decides the legal test.** Identify the scale first: it fixes the permissible summary statistic and the permissible test before you look at anything else in the analysis.

### 9.2 Nominal: named categories with no order

**Labels that differ in kind, not in amount.** **Nominal** (categorical) data have values that are qualitatively different from one another but carry no inherent order (Companion to Psychiatric Studies). The textbook example is diagnosis: schizophrenia, bipolar disorder and depression are distinct labels, but none is "more" than another, and no arithmetic on the labels is meaningful. Blood group, marital status, sex and the presence or absence of a symptom are all nominal. When a nominal variable has exactly two categories (alive or dead, case or non-case) it is called **dichotomous** or binary. The only legitimate summary is a count or a **proportion**; the only centre is the mode; and comparisons between groups use the chi-square test (or Fisher's exact test when expected cell counts are small).

### 9.3 Ordinal: ordered categories with unequal gaps

**Rank is real, but the spacing is not.** **Ordinal** data have a logical order in their values, but the differences between adjacent values are not necessarily equal and there is no true zero point (Companion to Psychiatric Studies). A single Likert item (strongly disagree, disagree, neutral, agree, strongly agree), tumour stage (I to IV), social class I to V, and a ward observation level all rank the units without promising that the step from stage I to stage II equals the step from stage III to stage IV. Because the gaps are not guaranteed equal, the mean is not strictly meaningful; the correct centre is the **median** with the interquartile range for spread, and group comparisons use **non-parametric** (distribution-free) tests such as the Mann-Whitney U test or the Wilcoxon signed-rank test.

TRAP **Treating an ordinal Likert score as interval without care.** Averaging a single Likert item assumes the psychological distance from "agree" to "strongly agree" equals that from "neutral" to "agree", which is not established; report the median, or justify interval treatment only for a summed multi-item scale.

### 9.4 Interval: equal gaps, no true zero

**Differences are meaningful; ratios are not.** **Interval** data have equal differences between values, so the distance between 1 and 2 is the same as the distance between 33 and 34, but the zero is arbitrary rather than absolute (Companion to Psychiatric Studies). The canonical example is temperature in Celsius: 20 degrees is 10 degrees warmer than 10 degrees, yet 20 degrees is not "twice as hot" as 10 degrees, and 0 degrees Celsius does not mean an absence of temperature. Because the zero is not true, sums and differences are valid but ratios are not. Interval data support the **mean** and standard deviation and, given the distributional assumptions, **parametric** tests. Calendar year and IQ score are usually treated as interval.

### 9.5 Ratio: equal gaps and a true zero

**The full arithmetic, ratios included.** **Ratio** data share the equal-interval property of interval data and add a true, non-arbitrary zero that marks a complete absence of the quantity (Companion to Psychiatric Studies). Weight, height, age, and number of admissions are ratio: zero admissions genuinely means none, and four admissions are meaningfully twice two. This true zero is what licenses ratios, so "twice as heavy" is a valid statement whereas "twice as hot in Celsius" is not. In practice the interval versus ratio distinction rarely changes the analysis: both are numerical, both are summarised by the mean and standard deviation, and both are analysed with the same parametric tests. That is why many texts pool them as "interval or ratio".

### 9.6 The categorical versus numerical split, and discrete versus continuous

**The higher-level partition that maps onto the four scales.** The four scales collapse into two families. **Categorical** data (nominal and ordinal) place units into classes; numerical data (interval and ratio) attach a genuine number. Within numerical data a second cut matters: **discrete** data can take only separate, usually whole-number, values (number of admissions, number of previous episodes, family size, a count of anything), whereas **continuous** data can in principle take any value on a continuum, limited only by measurement precision (weight, height, serum lithium concentration, reaction time). A count is discrete even though it is ratio-scaled; a measured quantity is continuous. The practical payoff is that discrete-count and continuous data, being numerical, are described by the mean and analysed by parametric methods, while categorical data are described by proportions and analysed by chi-square and its relatives (Companion to Psychiatric Studies).

**Categorical**
Data sorted into classes; comprises the nominal and ordinal scales; summarised by counts and proportions.

**Numerical**
Data carrying a true number; comprises the interval and ratio scales; summarised by the mean (or median if skewed).

**Discrete**
Numerical data limited to separate values, typically whole-number counts (number of admissions).

**Continuous**
Numerical data able to take any value on a continuum to the limit of measurement precision (weight, concentration).

### 9.7 Variable roles: independent and dependent, exposure and outcome, confounder

**The scale says what a variable is; its role says what it does.** A single **variable** can occupy different roles in a study, and naming the role is separate from naming the scale. The independent variable is the presumed cause, the one manipulated or classified on (treatment arm, drug dose); the dependent variable is the measured response that is expected to change with it (symptom score, remission). In epidemiology the same idea is renamed: the exposure is the putative causal factor (a risk factor, a treatment) and the **outcome** is the event or state under study (relapse, mortality). A confounder is a third variable independently associated with both the

exposure and the outcome, and not on the causal pathway between them, which distorts the crude exposure-outcome association if left unadjusted (age confounding an association between comorbidity and mortality is the standard example). Roles and scales are orthogonal: an outcome can be nominal (relapsed or not), ordinal (much improved to much worse), or ratio (days to relapse), and the outcome's scale, not its role, still dictates the test.

**WORKED EXAMPLE**

A trial randomises patients to drug A or drug B and, at 12 weeks, records (i) remission, yes or no, and (ii) the Hamilton depression score. Here the treatment arm is the **independent variable** (nominal, two categories); remission is a **nominal** dependent variable, so groups are compared with a chi-square test on proportions; the Hamilton score is treated as an **interval** dependent variable, so groups are compared with an independent-samples t-test on means. One study, two outcomes, two different tests, chosen entirely by the scale of each outcome. Had the outcome instead been an **ordinal** global-improvement rating, the correct test would switch to the non-parametric Mann-Whitney U test.

### 9.8 How the scale drives the summary statistic and the test

**One scale, one centre, one test family.** The chain is direct (Companion to Psychiatric Studies; Bland 2015). **Nominal** data are summarised by proportions and compared with the chi-square test; there is no meaningful measure of central tendency beyond the mode, and no measure of spread. **Ordinal** data are summarised by the median and interquartile range and compared with non-parametric tests. **Interval** and **ratio** data, if reasonably normally distributed, are summarised by the mean and standard deviation and compared with parametric tests (t-test, ANOVA). The scale you can always safely drop *down* to a lower one (treat interval data as ordinal by ranking, or as nominal by categorising) at the cost of throwing away information; you may never lawfully move a variable *up* a scale it does not possess. This is the reason ordinal-treated-as-interval is a genuine error while interval-treated-as-ordinal is merely wasteful.

| SCALE | EXAMPLE | CENTRE STATISTIC | TEST FAMILY |
|---|---|---|---|
| **Nominal** | Diagnosis; blood group; alive or dead | Mode; report as a proportion | **Chi-square (Fisher's exact if cells small)** |
| **Ordinal** | Likert item; tumour stage I to IV; social class | Median, interquartile range | **Non-parametric (Mann-Whitney, Wilcoxon)** |
| **Interval** | Temperature in Celsius; IQ; calendar year | Mean, standard deviation | **Parametric (t-test, ANOVA)** |
| **Ratio** | Weight; age; number of admissions | Mean, standard deviation | **Parametric (t-test, ANOVA)** |

The four scales of measurement read across a worked example, the honest measure of central tendency, and the family of tests each scale permits; the hit column is the test-family discriminator that the scale forces on you.

### 9.9 The statistical-test selector: data type by groups by pairing

**Once the scale is fixed, three further questions name the test.** Beyond the scale, the correct test for a difference depends on how many groups are compared and whether the observations are

paired (the same subjects measured twice, or matched) or unpaired (independent groups). The selector below is the version examined most heavily: read the data type down and the group structure across (Companion to Psychiatric Studies; Altman 1991). Note the clean parallelism, every parametric test on the right has a non-parametric partner on the left, and every categorical row uses a chi-square-family test.

| COMPARISON | CATEGORICAL (NOMINAL) | ORDINAL / NON-NORMAL | INTERVAL OR RATIO (NORMAL) |
|---|---|---|---|
| **Two groups, unpaired** | **Chi-square test** | Mann-Whitney U test | Independent-samples t-test |
| **Two groups, paired** | **McNemar test** | Wilcoxon signed-rank test | Paired t-test |
| **Three or more, unpaired** | **Chi-square test** | Kruskal-Wallis test | One-way ANOVA |
| **Three or more, paired** | **Cochran's Q test** | Friedman test | Repeated-measures ANOVA |
| **Association of two variables** | **Chi-square / phi** | Spearman rank correlation | Pearson correlation |

The statistical-test selector: data type down, number of groups and pairing across; the hit column is the categorical-data route, and each non-parametric test in the ordinal column is the distribution-free partner of the parametric test to its right.

**HOLD THIS**

The measurement scale is the first and decisive choice: nominal gives proportions and chi-square, ordinal gives the median and non-parametric tests, interval and ratio give the mean and parametric tests, and you may drop a variable down a scale but never push it up one it does not have.

**COMMON TRAP**

The commonest measurement error in psychiatry research is running a t-test or reporting a mean on data that are only **ordinal**, above all on single **Likert** items and clinician rating scales, as if the equal-interval property held. A count that looks numerical can also mislead: "number of admissions" is **discrete** and often heavily skewed, so its mean can be a poor centre and a non-parametric test or a count model is safer than a naive t-test.

**PEARL**

Interval and ratio almost never need to be distinguished in practice: both are numerical, both take the mean and standard deviation, and both are analysed by the same parametric tests, which is why textbooks routinely pool them as "interval or ratio" in their test tables (Companion to Psychiatric Studies). The examinable distinction is one rung lower, categorical (nominal, ordinal) versus numerical (interval, ratio), because that boundary is exactly where parametric analysis becomes permissible.

MNEMONIC

**NOIR, in rising order of information: Nominal, Ordinal, Interval, Ratio.**

Read left to right you climb from named categories (Nominal) to ranked categories (Ordinal) to equal gaps without a true zero (Interval) to equal gaps with a true zero (Ratio). Each step adds one permitted operation: equality, then order, then addition, then ratios. The categorical break sits between Ordinal and Interval, which is the parametric threshold.

GOOD TO REMEMBER

- **Nominal:** named categories with no order (diagnosis, blood group); use it for qualitative labels; the summary is a proportion and the test is the chi-square test (Fisher's exact when expected cell counts are small).
- **Ordinal:** ordered categories with unequal, unguaranteed gaps and no true zero (Likert item, tumour stage); use it for ranked data; the centre is the median with the interquartile range and the tests are non-parametric (Mann-Whitney, Wilcoxon).
- **Interval:** equal gaps but an arbitrary zero (temperature in Celsius, IQ); use it for numerical data with no true zero; the summary is the mean and standard deviation and the tests are parametric (t-test, ANOVA); differences are valid but ratios are not.
- **Ratio:** equal gaps and a true zero marking absence of the quantity (weight, age, number of admissions); use it for numerical data where "twice as much" is meaningful; summary and tests are the same as interval (mean, SD, parametric), with ratios now also valid.
- **Discrete versus continuous:** discrete numerical data take separate whole-number values (counts of admissions); continuous numerical data take any value on a continuum (weight, concentration); both are numerical, but a skewed count may need a non-parametric test despite being ratio-scaled.
- **Confounder:** a variable independently linked to both exposure and outcome and not on the causal pathway; recognise it whenever a crude association may be distorted by a third factor; it is controlled by stratification, matching, restriction or multivariable adjustment.

## 10 · Describing data: central tendency, dispersion and the normal curve

**Every dataset must first be described before it can be tested.** Descriptive statistics answer three questions about a set of numbers: where is its centre, how spread out is it, and what shape does it take. **Central tendency** gives the typical value, **dispersion** gives the scatter around that value, and **shape** (symmetry and peakedness) decides which summary is honest and which inferential test is legitimate. The examiner tests this as a chain: read the data type and shape, pick the correct measure of centre and spread, and only then reach for a test (Companion to Psychiatric Studies; Bland 2015; Altman 1991).

The keystone is the **normal distribution**, the bell-shaped curve on which most parametric statistics rest. Two properties make it examinable: it is fully specified by just its mean and standard deviation, and a fixed, memorisable fraction of the data falls within one, two and three standard deviations of the mean. Master that and a large block of biostatistics becomes arithmetic (Kaplan & Sadock 2024). The rest of this topic builds the vocabulary of centre, spread and shape, then closes with the probability ideas that feed formal inference.

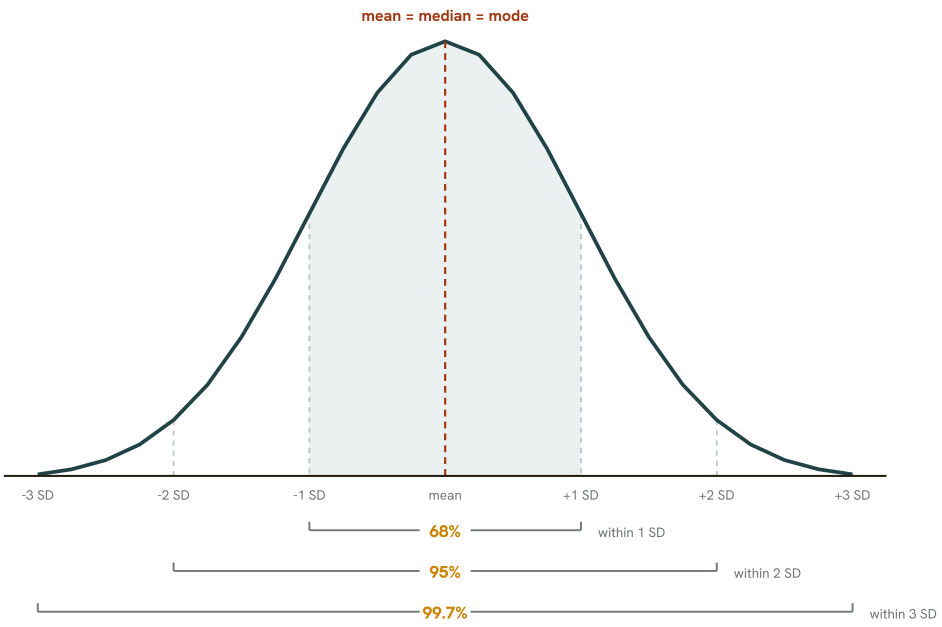



**Fig 7.4** The normal curve and the empirical rule. In a normal distribution the mean, median and mode coincide at the peak and the spread is measured in standard deviations. The empirical rule fixes the proportions worth memorising: about 68 percent of observations lie within one SD of the mean, about 95 percent within two, and about 99.7 percent within three. The 95 percent band (more precisely 1.96 SD) is the basis of most confidence intervals and reference ranges.

## 10.1 Central tendency: the three averages

**Central tendency is the single value that best represents a distribution.** There are three measures, and the correct one depends on the data type and the shape (Companion to Psychiatric Studies). The mean (the arithmetic average, sum of values divided by n) uses every observation and is the natural summary for symmetric, interval or ratio data. The median (the middle value when the data are ranked, the 50th percentile) splits the data into two equal halves and is the summary of choice for skewed data and for ordinal data. The mode (the most frequently occurring value) is the only measure that can be used for nominal (categorical) data, and it is the one that can be read off a bar chart. A distribution can have one mode (unimodal), two (bimodal), or none.

RULE **Report the median for skewed data.** When a distribution is skewed or the data are ordinal, the median is the honest centre; the mean is right only for symmetric interval or ratio data, and the mode is for nominal data.

## 10.2 Choosing the centre: mean vs median vs mode

**Match the measure to the data.** The mean is efficient but not robust: a single extreme value (an outlier) or a long tail pulls it away from the bulk of the data. The median ignores the actual magnitude of extreme values and cares only about rank, which is exactly why it is robust and why it is preferred for incomes, lengths of stay, symptom counts and any right-skewed clinical variable. The mode is crude for continuous data but indispensable for nominal categories such as

diagnosis or blood group, where "average" is otherwise meaningless. The selector table at 10.10 formalises this against data type.

**WORKED EXAMPLE**

Length of stay for seven patients: 3, 4, 4, 5, 6, 8, 40 days. The mean is (3+4+4+5+6+8+40)/7 = 70/7 = **10** days, which no typical patient experienced. The ranked middle value, the median, is **5** days and describes the group far better. The mode is **4** days (the commonest value). The single long-stay outlier at 40 days has dragged the mean upward, the classic reason to report the median for a right-skewed variable.

### 10.3 Dispersion: range and interquartile range

**Dispersion measures how widely the observations scatter around the centre.** The simplest is the range (the difference between the largest and smallest value, or the pair itself). It is quick but fragile, because it depends entirely on the two most extreme observations and grows with sample size. Its robust cousin is the interquartile range, the IQR, which is the distance between the 25th percentile (first quartile, Q1) and the 75th percentile (third quartile, Q3), so it spans the central 50 percent of the data. Because the IQR discards the extreme quarters at each end, it is the natural partner of the median for skewed data, and it is what the box of a box-and-whisker plot draws.

### 10.4 Dispersion: variance and standard deviation

**The variance and standard deviation are the dispersion measures that pair with the mean.** The variance is the average of the squared deviations of each value from the mean (Companion to Psychiatric Studies). Squaring removes the sign so that deviations do not cancel, but it leaves the answer in squared units (for example, days squared), which is hard to interpret. Taking the square root of the variance returns to the original units and gives the standard deviation (SD), the single most used measure of spread for symmetric data. A large SD means the data are widely scattered about the mean; a small SD means they cluster tightly. For sample data the variance divides the sum of squared deviations by n minus 1 (Bessel's correction) rather than n, which corrects the tendency of a sample to underestimate the population spread.

**HOLD THIS**

The mean pairs with the standard deviation (symmetric data); the median pairs with the interquartile range (skewed data). Never a mean with an IQR, never a median with an SD.

### 10.5 Standard deviation vs standard error of the mean

**SD describes the data; the standard error of the mean describes the estimate.** This is the single most confused pair in the examination (Companion to Psychiatric Studies). The standard deviation quantifies the spread of the individual observations in your sample and does not shrink as you collect more data (a bigger sample just estimates the true SD more precisely). The standard error of the mean (SEM) is the standard deviation of the sample mean itself, that is, how much the sample mean would vary from study to study; it equals the SD divided by the square root of n. Because n is in the denominator, the SEM falls as the sample grows, so it is always smaller than

the SD. Use the SD to describe how variable the patients are; use the SEM (and the confidence interval built from it) to say how precisely you have pinned down the mean.

TRAP **SEM is not a measure of data spread.** Authors quote the small SEM instead of the SD to make data look tidy; the SEM only reflects precision of the mean and always shrinks with n. To describe the scatter of patients, report the SD.

### 10.6 The normal distribution and the empirical rule

**The normal distribution is the symmetric, bell-shaped curve where mean = median = mode.** Also called the gaussian distribution, it is unimodal and perfectly symmetric about its centre, so the three measures of central tendency coincide at the single peak (Companion to Psychiatric Studies; Kaplan & Sadock 2024). It is completely defined by two parameters, the mean (which locates the peak) and the standard deviation (which sets the width). Its key examinable property is the empirical rule (the 68 to 95 to 99.7 rule): for any normal distribution, about **68** percent of observations lie within one SD of the mean, about **95** percent within two SDs, and about **99.7** percent within three SDs. These bands, **68/95/99.7**, are the most reusable numbers in descriptive statistics.

| 68/95/99.7 | 1.96 |
| --- | --- |
| PERCENT WITHIN 1, 2, 3 SD | SD FOR CENTRAL 95 PERCENT |

### 10.7 The standard normal and the z-score

**Standardising turns any normal value into a position on one universal curve.** The standard normal distribution is the special normal distribution with mean 0 and SD 1. Any raw value x from a normal distribution is converted to it by the z-score, z = (x minus mean) / SD, which expresses how many standard deviations the value sits above (positive z) or below (negative z) the mean. A z-score lets you compare values from different scales and read off the proportion of the distribution beyond any point from a single z-table. The empirical rule is just the z-table at round numbers: z = 1, 2 and 3 cut off the 68, 95 and 99.7 percent bands. More precisely, the central 95 percent of a normal distribution lies within **1.96** SDs of the mean (not exactly 2), and 99 percent within 2.58 SDs; the 1.96 value is the one to memorise because it reappears as the multiplier in the 95 percent confidence interval.

WORKED EXAMPLE

IQ is scaled to a normal distribution with mean **100** and SD **15**. A score of 130 has z = (130 minus 100)/15 = 30/15 = **2**, so it sits exactly two SDs above the mean. By the empirical rule 95 percent of people lie within two SDs (between 70 and 130), leaving 5 percent outside; because the curve is symmetric, about 2.5 percent score above 130. Intellectual disability is conventionally set near two SDs below the mean, an IQ around 70, which is the same 2.5 percent tail at the lower end (DSM-5-TR 2022).

### 10.8 Shape: skew

**Skew is asymmetry, and it tells you where the mean has gone.** A skewed distribution has one tail longer than the other, and the direction is named for the tail, not the hump. In a right

(positive) skew the long tail points to the high values (as in length of stay, income, or reaction times), and the mean is dragged toward that tail so the order becomes mode less than median less than mean. In a left (negative) skew the long tail points to the low values and the order reverses to mean less than median less than mode. This is the mechanism behind the rule to report the median for skewed data: the mean chases the tail while the median stays near the bulk. A quick sign of right skew is a mean noticeably larger than the median.

**PEARL**

The mean is pulled toward the longer tail; the median resists it. So in right (positive) skew the mean exceeds the median, and in left (negative) skew the mean falls below the median. Comparing mean and median is a one-line skew detector.

### 10.9 Shape: kurtosis

**Kurtosis describes the peakedness and tail-weight of a distribution.** Where skew is about asymmetry, kurtosis is about how heavy the tails are and how sharp the peak is relative to a normal curve. A distribution with positive excess kurtosis (leptokurtic) is more sharply peaked with fatter tails, meaning more extreme outliers than a normal distribution predicts; one with negative excess kurtosis (platykurtic) is flatter with thinner tails. The normal distribution is the reference (mesokurtic, excess kurtosis 0). Kurtosis matters because fat tails break the assumptions of tests that presume normality, and heavy-tailed data are better summarised by the median and IQR.

### 10.10 Selector: which centre and spread for which data

**The measure of centre and its matched spread follow directly from data type and shape.** This is the practical payoff of the whole topic and a favourite one-mark question. Read the row, name the pair.

| DATA / SHAPE | BEST CENTRE | MATCHED SPREAD | WHY |
|---|---|---|---|
| **Symmetric interval/ratio** | **Mean** | Standard deviation | Uses all data; SD and mean fully describe a normal curve |
| **Skewed continuous** | **Median** | Interquartile range | Median and IQR ignore the outlying tail; robust |
| **Ordinal (ranked)** | **Median** | Interquartile range / range | Ranks have order but no equal intervals, so no true mean |
| **Nominal (categorical)** | **Mode** | (Counts / proportions) | Only the commonest category is meaningful; no arithmetic |

Descriptive selector: data type and shape fix the centre, which fixes the matched spread; the hit column is the answer the stem wants.

### 10.11 Probability: the bridge to inference

**Probability turns a described distribution into a basis for inference.** A probability is the long-run relative frequency of an outcome, a number from 0 (impossible) to 1 (certain), and the probabilities of all mutually exclusive outcomes sum to 1 (Companion to Psychiatric Studies). Two ideas carry forward. First, for a continuous variable, area under the curve equals probability, which is exactly what the empirical rule computes: the 95 percent within 1.96 SDs is the area under the central slab of the normal curve. Second, two events are independent when the occurrence of one does not change the probability of the other, the assumption behind treating patients as independent observations. These feed the sampling distribution of the mean, the standard error, confidence intervals and hypothesis tests, where descriptive statistics hand over to formal inference.

MNEMONIC

**Mean Moves, Median Middles**

In a skewed distribution the MEAN MOVES toward the long tail while the MEDIAN holds the MIDDLE. So report the median (and IQR) whenever the data are skewed, and reserve the mean (and SD) for symmetric data.

GOOD TO REMEMBER

- **Mean:** arithmetic average using all values; for symmetric interval/ratio data; = sum of values / n; pairs with the SD.
- **Median:** the middle (50th percentile) ranked value; for skewed or ordinal data; robust to outliers; pairs with the IQR.
- **Mode:** the most frequent value; the only centre valid for nominal data; read off a bar chart.
- **Range:** largest minus smallest value; crude, driven by the two extremes and grows with n.
- **Interquartile range:** Q3 minus Q1 (75th minus 25th percentile); spans the central 50 percent; robust partner of the median.
- **Variance:** mean of squared deviations from the mean; in squared units; sample version divides by n minus 1.
- **Standard deviation:** square root of the variance, back in original units; the spread that pairs with the mean.
- **Standard error of the mean:** SD of the sample mean = SD / square root of n; a precision measure that shrinks with n, always smaller than the SD.
- **Normal distribution (gaussian):** symmetric bell curve, mean = median = mode; fixed by mean and SD; empirical rule 68/95/99.7 percent within 1, 2, 3 SDs.
- **Standard normal / z-score:** normal with mean 0, SD 1; z = (x minus mean) / SD = number of SDs from the mean; central 95 percent lies within 1.96 SDs.
- **Skew:** asymmetry named for the long tail; right/positive tail drags the mean above the median, left/negative below it.
- **Kurtosis:** peakedness and tail-weight vs normal; leptokurtic = peaked, fat tails; platykurtic = flat, thin tails.

## 11 · Tests of significance: parametric and non-parametric

A significance test asks a single question: is the difference or association we observed likely to be real, or is it plausibly the play of chance? Every test we cover here re-frames the clinical question

as a **null hypothesis** of no difference, computes a test statistic, and returns a p-value, the probability of a result this extreme (or more) if the null were true (Johnstone et al 2010, Bland 2015). The examiner rarely asks you to compute anything by hand. What they test is whether you can **choose the right test** for the data in front of you, and whether you know why the wrong choice is wrong. That choice turns on three things: the scale of your outcome (nominal, ordinal, interval or ratio), how many groups you are comparing, and whether the observations are paired.

The great divide is between parametric tests, which assume the data follow a known distribution (usually normal) and estimate population parameters like the mean and variance, and non-parametric tests (also called distribution-free tests), which make no such assumption and work on ranks (Altman 1991, Kaplan & Sadock 2024). Parametric tests are more powerful when their assumptions hold; the non-parametric twins are the safety net for ordinal or non-normal data.

### 11.1 The three parametric assumptions

**What must be true for a parametric test.** The correct choice of a parametric test rests on a short list of assumptions (Johnstone et al 2010): the outcome is **interval or ratio** data (continuous, with meaningful equal intervals); each observation is **independent** of the others; the data are approximately **normally distributed**; and the **variance of each group is approximately equal** (homoscedasticity). Meet these and you may use the powerful mean-based tests that use every data value. Fail them badly and you switch to a non-parametric rank-based test, or transform the data (e.g. by taking the log or square root) to restore normality first.

**Robustness.** A useful caveat: parametric tests are relatively robust to minor departures from normality, contrary to the common misperception that perfect normality is the overriding requirement (Johnstone et al 2010). With a large sample the central limit theorem means sample means tend to normality even when the raw data do not, so a large-sample t-test remains safe even when some assumptions are mildly violated.

RULE **Scale plus groups plus pairing.** Match the test to the data scale, the number of groups, and whether the observations are paired: those three facts pick the test.

### 11.2 The student t-test: one-sample, two-sample, and paired

**The parametric workhorse for one or two groups.** The student t-test (use the t-distribution) is one of the most quoted statistics in medicine (Johnstone et al 2010). It comes in three flavours matched to the design:

#### One-sample t-test

Compares a single sample mean against a known or hypothesised population mean (e.g. mean age of a dementia cohort vs mean age of all elderly inpatients).

#### Independent two-sample t-test

Compares the means of two independent groups; needs continuous or interval data, approximate normality, and roughly equal variances.

#### Paired t (paired t-test)

For matched data: the same subjects measured twice (before/after) or individually matched pairs. It analyses the within-pair differences, not the raw group means.

**The statistic.** In essence, t = observed difference in means / standard error of that difference, with **n1 + n2 – 2** degrees of freedom for the two-sample test. t grows (and significance becomes more likely) as the difference in means widens or as the standard error shrinks with larger samples (Johnstone et al 2010). Where the equal-variance assumption fails there is a modified two-sample t-test (Welch). Always state t and its degrees of freedom.

### 11.3 ANOVA and the F-test: three or more groups

**Why you cannot just run more t-tests.** When there are three or more groups you test the null hypothesis that all group means are equal using the analysis of variance (ANOVA). Counter-intuitively, ANOVA answers a question about means by comparing variances (Johnstone et al 2010). It partitions the total variation into the between-groups variance (how far the group means sit from the grand mean) and the within-groups (residual) variance (scatter inside each group), and takes their ratio as the test statistic F-test (the F ratio):

F = variation between groups / variation within groups = MS(between) / MS(residual). If the groups share one population mean, F is near 1; if the group means truly differ, the between-groups term swells and F rises. A significant F says at least one group differs, but not which one.

TRAP **Do not t-test three groups.** Running a t-test on every pair of three groups inflates the type I error (each test carries its own 5% false-positive risk). Use ANOVA for three or more groups.

**Extensions.** One-way ANOVA handles a single grouping factor; factorial ANOVA handles several factors at once and reveals interactions; repeated-measures ANOVA is the paired extension for the same subjects measured on several occasions; ANCOVA adjusts the group comparison for a confounding covariate such as age or premorbid IQ (Johnstone et al 2010).

### 11.4 Post-hoc comparisons: Tukey and Bonferroni

**Finding which groups differ, safely.** A significant ANOVA licenses post-hoc pairwise comparisons to locate the difference, but every extra comparison is another chance for a false positive, so these tests build in a correction (Johnstone et al 2010). The Tukey Honestly Significant Difference and the Bonferroni adjustment are the two most-quoted. Bonferroni is the simplest and most conservative: to keep the overall false-positive rate at 0.05 across n tests, each individual test must clear a threshold of p = **0.05** / number of tests. For two comparisons that is roughly 0.025; for many comparisons it becomes very stringent, so critics note it can over-correct and inflate the type II error (Perneger 1998).

### 11.5 Mann-Whitney U: the non-parametric two independent groups test

**The rank-based twin of the independent t-test.** When the outcome is ordinal, or interval data that badly break normality, and you are comparing two independent groups, use the mann-whitney U test (Johnstone et al 2010). It ranks all observations across both groups combined and compares the rank sums, so it makes no distributional assumption. It assumes only that the two

groups are independent, and because it can detect differences in spread as well as in location, report each group's median alongside a note on skew (e.g. a box plot). Always state U.

### 11.6 Wilcoxon signed-rank: the non-parametric paired test

**The rank-based twin of the paired t.** Where measurements are paired (two measurements from the same individual, or matched pairs) and the data are ordinal or non-normal, the wilcoxon signed-rank (matched-pairs) test replaces the paired t-test (Johnstone et al 2010). It ranks the within-pair differences by magnitude, attaches the sign, and tests whether the signed ranks balance around zero. Pair Mann-Whitney with independent groups and Wilcoxon with paired data: swapping them is the classic slip.

### 11.7 Kruskal-Wallis: the non-parametric ANOVA

**The rank-based twin of one-way ANOVA.** For three or more independent groups on ordinal or non-normal data, the kruskal-wallis test is the non-parametric equivalent of one-way ANOVA (Johnstone et al 2010). It ranks all observations and tests whether the mean ranks differ across groups; a significant result, like a significant F, says at least one group differs and is followed by post-hoc rank comparisons. Its paired counterpart (repeated measures on the same subjects) is the Friedman test.

### 11.8 Chi-square: association for categorical data

**Comparing proportions, not means.** For categorical (nominal) outcomes the chi-square (chi square) test of association is the key test (Johnstone et al 2010). It is distribution-free, and it asks whether the proportion with a characteristic differs across two or more independent groups by comparing observed cell counts with the counts expected under the null. The statistic is chi-square = sum of (O − E)squared / E, where O and E are observed and expected frequencies. Two rules the examiner loves: the cells must contain absolute counts, never percentages (percentages falsely inflate the sample size); and you must state chi-square with its degrees of freedom. The paired analogue for a 2x2 table (same subjects, before/after) is the McNemar test.

### 11.9 Fisher exact: the small-cell rescue

**When the chi-square approximation breaks.** The chi-square distribution is an approximation that fails when expected cell counts are small. The classic rule: when any expected frequency falls below **5**, use the fisher exact test (or Yates's continuity correction) instead of the ordinary chi-square (Johnstone et al 2010). Fisher's exact computes the exact probability of the observed table and is more conservative, so it is less likely to reject the null wrongly when numbers are tiny.

| 0.05 | < 5 | n−1 |
|---|---|---|
| ALPHA THRESHOLD | EXPECTED CELL TO FISHER | DEGREES OF FREEDOM IDEA |

### 11.10 The test-map: pairing each parametric test with its non-parametric twin

**The one table to hold.** Read across data type, number of groups, and pairing to land on the correct test. This is the selector the examiner is really testing (Johnstone et al 2010, Altman 1991).

| GROUPS & PAIRING | CATEGORICAL (NOMINAL) | ORDINAL / NON-NORMAL | INTERVAL OR RATIO (PARAMETRIC) |
|---|---|---|---|
| **One sample vs known value** | Chi-square goodness-of-fit | Sign test | **One-sample t-test** |
| **Two groups, independent (unpaired)** | Chi-square test of association | Mann-Whitney U | **Independent two-sample t-test** |
| **Two groups, paired / matched** | McNemar test | Wilcoxon signed-rank | **Paired t-test** |
| **Three or more groups, independent** | Chi-square test | Kruskal-Wallis | **ANOVA (one-way), F-test** |
| **Three or more groups, repeated / paired** | Cochran's Q | Friedman test | **Repeated-measures ANOVA** |
| **Association between two variables** | Chi-square / phi coefficient | Spearman rank correlation | **Pearson correlation** |

Every parametric test in the right-hand column has a rank-based non-parametric twin to its left; small-cell categorical data drop to the Fisher exact test.

## 11.11 Worked reads: a chi-square 2x2 and an ANOVA F

WORKED EXAMPLE

**Chi-square, 2x2.** Sex distribution is compared between 60 people with schizophrenia and 60 healthy controls (Johnstone et al 2010). Expected counts under no difference are 30 per cell. Summing (O − E)squared / E across the four cells gives chi-square = 1.2 on 1 degree of freedom, p = 0.27. Because 0.27 far exceeds alpha 0.05, the null of no difference in the population proportion cannot be rejected: the sexes are not shown to differ.

**ANOVA F.** Ages of three diagnostic groups of four patients are compared (Johnstone et al 2010). The between-groups sum of squares is 1711.5 (2 degrees of freedom) and the within-groups (residual) sum of squares is 648.5 (9 degrees of freedom). Dividing each by its degrees of freedom gives mean squares of 855.75 and 72.1, so F = 855.75 / 72.1 = 11.9, p = 0.003. Since 0.003 is below 0.05 we reject the null: at least one group mean differs, and post-hoc tests would then locate which.

COMMON TRAP

Feeding percentages into a chi-square test instead of raw counts falsely inflates the sample size and manufactures spurious significance; chi-square needs absolute frequencies. Equally, treating an ordinal scale (e.g. a 1-to-6 social-class rank) as interval data and running a t-test is a scale error: use Mann-Whitney or Kruskal-Wallis instead.

MNEMONIC

**MWWK guards the non-parametric gate: Mann, Wilcoxon, Whitney's cousin, Kruskal.**

Two independent groups to **Mann**-Whitney; two paired to **Wilcoxon**; three or more independent to **Kruskal**-Wallis. Their parametric partners are the independent t-test, the paired t-test, and ANOVA in the same order.

HOLD THIS

Match the test to data scale x number of groups x paired?: t-test twin is Mann-Whitney (2 independent) or Wilcoxon (paired), ANOVA twin is Kruskal-Wallis (3+), and categorical goes to chi-square, dropping to Fisher exact when any expected cell is under 5.

GOOD TO REMEMBER

- **Student t-test:** parametric test of means for one or two groups; use when data are interval/ratio, roughly normal, with equal variance; two-sample t has n1 + n2 − 2 degrees of freedom.
- **Paired t-test:** parametric test for matched/before-after data; use when the same subjects are measured twice; it tests the mean of the within-pair differences.
- **ANOVA (analysis of variance):** parametric test for three or more group means; use to avoid multiple t-tests; its statistic is F = MS(between) / MS(residual).
- **F-test:** the ratio of between-groups to within-groups variance inside ANOVA; use to test whether group means are equal; F near 1 supports the null.
- **Post-hoc (Tukey, Bonferroni):** follow a significant ANOVA to find which groups differ; use to control the family-wise error; Bonferroni threshold = 0.05 / number of tests.
- **Mann-Whitney U:** non-parametric test for two independent groups on ordinal/non-normal data; use as the t-test's rank-based twin; report the group medians.
- **Wilcoxon signed-rank:** non-parametric test for paired ordinal/non-normal data; use as the paired-t twin; it ranks the signed within-pair differences.
- **Kruskal-Wallis:** non-parametric test for three or more independent groups; use as the one-way ANOVA twin; the paired counterpart is the Friedman test.
- **Chi-square (chi square) test of association:** non-parametric test comparing proportions across categorical groups; use on absolute counts, never percentages; statistic = sum of (O − E)squared / E, stated with its degrees of freedom.
- **Fisher exact test:** the small-cell rescue for a 2x2 table; use when any expected cell frequency is under 5; it gives the exact, more conservative probability.

For test-selection detail and worked derivations see Altman (1991) and Bland (2015). Reliability and validity of a psychometric instrument itself (as distinct from the reliability of a study) are covered in the Psychometry, Assessment and Descriptive Psychopathology notes; the disorder-specific applications of these tests appear in the clinical chapters of Paper II.

## 12 · Correlation, regression and factor analysis

**Once two or more variables are in play, three questions dominate the stems: do they move together, can one predict another, and do many items collapse into a few?** This topic covers the three workhorse techniques of association and prediction: **correlation** quantifies how strongly two variables move together, **regression** models and predicts one variable from others

(and adjusts for confounders), and **factor analysis** reduces many correlated items to a few underlying latent dimensions (Companion to Psychiatric Studies; Bland 2015; Altman 1991). The examinable spine is simple: correlation measures, regression predicts and adjusts, factor analysis simplifies, and none of the three, on its own, proves cause.

The single most tested idea across the whole cluster is the divorce of statistical association from causation: a strong correlation, a steep regression slope, or a high coefficient of determination each describes the data at hand but licenses no causal claim without a design that supports one (Tasman 2015; Kaplan & Sadock 2024). Hold that caveat and the traps in this area mostly disarm themselves.

### 12.1 Correlation: measuring linear association between two variables

**A single number for how tightly two variables track.** A correlation coefficient summarises the strength and direction of the linear relationship between two variables in one number that runs from **-1** through **0** to **+1** (Bland 2015; Altman 1991). A coefficient of **+1** is a perfect positive linear relationship (as one rises the other rises exactly in step), **-1** is a perfect negative linear relationship, and **0** means no linear relationship at all. The sign gives the direction and the magnitude gives the strength; a value near zero does not exclude a strong non-linear (for example U-shaped) relationship, because the coefficient only captures the linear component. Correlation is symmetric: the correlation of x with y equals the correlation of y with x, which is exactly why it can never, by itself, say which variable is cause and which is effect.

### 12.2 Pearson r: the coefficient for two continuous, normal variables

**The parametric correlation, for continuous normally distributed data.** The pearson product-moment correlation coefficient, written r, is the correlation used when both variables are continuous and approximately normally distributed and their relationship is linear (Bland 2015; Altman 1991). It is the default numerical-versus-numerical association measure and the parametric member of the correlation family, the natural partner of the t-test and ANOVA. Being parametric, it is sensitive to outliers and assumes a straight-line relationship; a single extreme point can inflate or collapse it, and a curved relationship can hide behind a near-zero r. Its square, r-squared, carries a direct interpretation covered below.

RULE **Match the correlation to the data.** Two continuous normal variables in a linear relationship take pearson r; ordinal or non-normal (or outlier-ridden) data take the spearman rank correlation.

### 12.3 Spearman rank correlation: for ordinal or non-normal data

**The non-parametric correlation, computed on ranks.** The spearman rank correlation coefficient (rho) is the distribution-free counterpart of pearson r: it is applied when the data are ordinal, when a continuous variable is markedly non-normal or skewed, or when outliers make a rank-based measure safer (Bland 2015; Companion to Psychiatric Studies). It is simply the pearson correlation computed on the ranks rather than the raw values, so it captures any monotonic (consistently increasing or decreasing) relationship, not only a strictly linear one, and it shares the same -1 to +1 range and the same sign convention. Because it works on ranks it is

robust to outliers and makes no normality assumption, at the modest cost of discarding the exact spacing of the values.

**WORKED EXAMPLE**

A study reports that clinician-rated illness severity (a five-point ordinal scale) and number of prior admissions (a skewed count) rise together, with a coefficient of **0.62**. Because one variable is ordinal and the other is a skewed count, pearson r is inappropriate and the **spearman** rank correlation is the correct choice; rho of 0.62 says the ranks move together moderately and positively. Note what it does not say: it does not establish that more admissions cause greater severity, nor the reverse, because a third factor (chronicity, poor adherence) could drive both. The coefficient describes the monotonic association and nothing more.

### 12.4 Correlation is not causation

**The caveat that governs the entire topic.** A correlation, however strong, cannot by itself be logically read as evidence that one variable causes the other; the non-equivalence of correlation and causation applies with full force (Tasman 2015). Two variables can correlate because one causes the other, because the other causes the one, because a third variable (a **confounder**) causes both, or by chance. The classic pattern is confounding: ice-cream sales correlate with drowning deaths, but summer heat drives both. Reverse causation and selection effects produce the same trap. Establishing causation needs a design that controls these alternatives (a randomised trial, or at minimum multivariable adjustment plus the Bradford Hill considerations), never a correlation coefficient on its own.

TRAP **Reading a strong r, slope, or r-squared as proof of cause.** Association is not causation; a confounder, reverse causation, or chance can generate any of them without a causal link between the two variables.

### 12.5 Linear regression: predicting a continuous outcome from a line

**From describing association to predicting a value.** Where correlation gives a symmetric strength-of-association number, linear regression fits an asymmetric predictive model: it estimates the straight line that best predicts a continuous outcome (the dependent variable, y) from one or more predictors (independent variables, x), by minimising the squared vertical distances of the points from the line (least squares) (Bland 2015; Altman 1991). The fitted line has the form y equals intercept plus slope times x. The **slope** (regression coefficient, b) is the heart of the output: it states how many units the outcome changes for each one-unit rise in the predictor, and its sign gives the direction. Unlike correlation, regression is directional, it names a predictor and an outcome, which is why the stem's phrasing ("predict", "for each unit increase") signals regression rather than correlation.

### 12.6 The coefficient of determination: r-squared as variance explained

**The proportion of the outcome's variability the model explains.** The coefficient of determination, written **r-squared**, is the square of the correlation coefficient and gives the proportion of the total variability in the outcome that is predictable from, or accounted for by,

the predictor(s) (Tasman 2015; Bland 2015). It runs from **0** to **1** (often quoted as a percentage): an r of **0.5** gives an r-squared of **0.25**, meaning **25%** of the variance in the outcome is explained by the predictor and the remaining 75% is unexplained. It is the standard effect-size and goodness-of-fit measure for a linear model. The critical exam point is that a high r-squared measures only how much variance the model captures; it says nothing about whether the relationship is causal.

| -1 to +1 | 0 to 1 | 0.25 |
|---|---|---|
| CORRELATION RANGE (R, RHO) | R-SQUARED RANGE | R-SQUARED WHEN R = 0.5 |

### 12.7 Logistic regression: predicting a binary outcome, yielding adjusted odds ratios

**The regression for yes/no outcomes.** When the outcome is binary (relapsed or not, dead or alive, case or non-case) a straight line is inappropriate, and logistic regression is used instead: it models the log-odds of the outcome as a linear function of the predictors (Bland 2015; Kaplan & Sadock 2024). Its coefficients, once exponentiated, yield odds ratios, and because several predictors sit in one model, each odds ratio is *adjusted* for the others in the model. An adjusted odds ratio therefore estimates the effect of one predictor on the odds of the outcome while holding the remaining predictors constant, which is precisely how confounding is handled in observational analyses. An odds ratio of **1** means no association; above 1 means increased odds; below 1 means decreased odds.

### 12.8 Multivariable regression: adjusting for confounders

**Putting several predictors in one model to isolate an effect.** The shared power of the regression family is that entering more than one predictor at once gives a **multivariable** model in which each coefficient is estimated with the others held constant, statistically adjusting the exposure-outcome association for measured confounders (Bland 2015; Altman 1991). This is the analytic answer to confounding introduced in the data-types and validity material: rather than adjusting by design (randomisation, matching, restriction), multivariable regression adjusts in the analysis. Linear regression does this for a continuous outcome, giving adjusted slopes; logistic regression does it for a binary outcome, giving adjusted odds ratios; and the parallel Cox model does it for time-to-event outcomes, giving adjusted hazard ratios. The unadjusted (crude) estimate becomes the adjusted estimate once confounders enter the model.

> **PEARL**
>
> The word that tells you which member of the regression family is meant is the *outcome*. A continuous outcome (symptom score, blood level) means **linear regression** and a slope; a binary outcome (relapse yes/no) means **logistic regression** and an adjusted odds ratio; a time-to-event outcome (time to relapse) means Cox regression and an adjusted hazard ratio. Same logic of adjustment, three outcome types, three models.

### 12.9 Factor analysis: reducing many correlated items to a few latent factors

**Data reduction for scale construction.** Factor analysis is a multivariate technique that identifies a small set of underlying, unobserved dimensions (latent factors) that explain the pattern of

intercorrelations among a larger set of measured items (Tasman 2015; Kaplan & Sadock 2024). Its starting point is the correlation matrix of all item pairs; from that matrix it extracts factors, each a weighted combination of items that hang together. The paradigm use in psychiatry is scale construction and validation: nine anxiety items might collapse to two factors (say, somatic and cognitive anxiety), giving a more parsimonious measure. The technique underpins the trait structure of personality questionnaires, where extraversion, neuroticism and psychoticism emerge as factors reproducible across cultures (Companion to Psychiatric Studies). Test-construction reliability and validity concepts sit alongside this in the Psychometry, Assessment and Descriptive Psychopathology notes.

### 12.10 Exploratory versus confirmatory factor analysis

**Discovering structure versus testing a hypothesised one.** Exploratory factor analysis is used when there is no firm prior model: it lets the data reveal how many factors exist and which items load on which, and is the tool for the first pass at a new scale (Tasman 2015). Confirmatory factor analysis is used later, when a specific factor structure is already hypothesised (from theory or from prior exploratory work): it tests how well a pre-specified model fits new data, reporting goodness-of-fit indices, and is a form of structural equation modelling. The examinable contrast is direction of reasoning: exploratory factor analysis generates the structure, confirmatory factor analysis tests a structure that is already specified.

### 12.11 Eigenvalues, the Kaiser criterion, loadings and the scree plot

**The machinery for deciding how many factors to keep and what they mean.** An eigenvalue is the amount of total variance in the item set that a given factor accounts for; larger eigenvalues mark more important factors. The commonest retention rule is the **Kaiser criterion**: retain only factors with an eigenvalue greater than **1**, on the logic that a retained factor should explain more variance than a single original item (Tasman 2015). The scree plot graphs eigenvalues in descending order; the factors above the "elbow" (the point where the curve levels into scree) are retained. A factor loading is the correlation between an item and a factor, and is how factors are interpreted: an item with a high loading defines that factor, and rotation (commonly varimax) is applied to sharpen the pattern so that each item loads highly on one factor and negligibly on the rest. Determining the number of factors is a matter of judgement, not an automatic output.

### 12.12 Principal component analysis and its relation to factor analysis

**The close cousin, and how it differs.** Principal component analysis is the related and often-conflated data-reduction method that repackages a set of correlated variables into a smaller set of uncorrelated components ordered by the variance each explains (Tasman 2015). In the factor-analysis workflow, using a communality estimate of 100% (retaining 1.0 on the diagonal of the correlation matrix) makes the extraction principal components. The conceptual difference is what each seeks: principal component analysis is purely descriptive and maximises variance explained, treating components as summaries of the observed variables, whereas true factor analysis (for example principal-factors extraction) posits latent factors that *cause* the observed correlations and separates common variance from item-specific and error variance. In practice they often give similar solutions, and principal component analysis is frequently the default first step.

**RULE** **Correlation quantifies association; regression predicts and adjusts.** Reach for a correlation coefficient to measure how two variables move together, and for regression to predict an outcome and to adjust for confounders.

| TECHNIQUE | WHAT IT DOES | OUTCOME / DATA | KEY OUTPUT NUMBER |
|---|---|---|---|
| **Pearson r** | Linear association, two variables | Both continuous, normal | **r, from -1 to +1** |
| **Spearman rho** | Monotonic association on ranks | Ordinal or non-normal | **rho, from -1 to +1** |
| **Linear regression** | Predict a continuous outcome; adjust | Continuous outcome | **Slope (b); r-squared** |
| **Logistic regression** | Predict a binary outcome; adjust | Binary outcome | **Adjusted odds ratio** |
| **Factor analysis** | Reduce items to latent factors | Many correlated items | **Eigenvalue; factor loading** |

The association-and-prediction selector: read the technique down, and let the data type and the question (measure, predict, or reduce) fix the choice; the hit column is the signature output number each technique returns.

**HOLD THIS**

Correlation measures association (pearson r for continuous normal data, spearman rho for ordinal or non-normal, both running -1 to +1); regression predicts and adjusts (linear regression for a continuous outcome with a slope and r-squared as variance explained, logistic regression for a binary outcome with adjusted odds ratios); factor analysis reduces many correlated items to a few latent factors (eigenvalue greater than or equal to 1 by the Kaiser criterion, interpreted through factor loadings); and none of them, alone, proves causation.

**COMMON TRAP**

The flagship error is reading a high r-squared (or a strong r, or a steep slope) as proof of causation: the coefficient of determination only states how much of the outcome's variance the model explains, not why, and a confounder or reverse causation can produce a large r-squared with no causal link. A second, quieter trap is applying pearson r to ordinal or skewed data where the spearman rank correlation is required, or reading a near-zero pearson r as "no relationship" when the true relationship is strong but curved.

**MNEMONIC**

### Correlation Counts, Regression Reaches, Factor analysis Folds.

Correlation *counts* the association between two variables (one number, symmetric, no direction of cause). Regression *reaches* forward to predict an outcome and to adjust for confounders (directional, predictor to outcome). Factor analysis *folds* many correlated items into a few latent factors (data reduction for scale building).

**GOOD TO REMEMBER**

- **Pearson r:** the correlation coefficient for two continuous, approximately normal variables in a linear relationship; use it as the default numerical-versus-numerical association measure; r runs from -1 to +1 and its square is r-squared.
- **Spearman rank correlation (rho):** the non-parametric correlation for ordinal or non-normal data, computed on ranks; use it when data are ordinal, skewed, or outlier-prone; captures any monotonic relationship and also runs from -1 to +1.
- **Linear regression:** predicts a continuous outcome from predictors by a best-fit least-squares line; use it to predict a value or to adjust for confounders; the slope (b) gives the change in outcome per one-unit rise in the predictor, and r-squared gives the proportion of outcome variance explained.
- **Coefficient of determination (r-squared):** the square of the correlation, the fraction of outcome variance the model explains; use it as the goodness-of-fit and effect-size measure; runs 0 to 1, so r of 0.5 gives r-squared of 0.25, that is 25% of variance explained.
- **Logistic regression:** predicts a binary outcome by modelling its log-odds; use it when the outcome is yes/no; its exponentiated coefficients are adjusted odds ratios, an odds ratio of 1 meaning no association.
- **Multivariable regression:** any regression with several predictors entered together; use it to adjust the exposure-outcome estimate for measured confounders in the analysis; each coefficient is estimated with the others held constant, turning a crude estimate into an adjusted one.
- **Factor analysis:** reduces many correlated items to a few latent factors from their correlation matrix; use it for scale construction and validation; retain factors with an eigenvalue greater than or equal to 1 (Kaiser criterion) and interpret each factor through its factor loadings (the item-factor correlations).
- **Exploratory factor analysis versus confirmatory factor analysis:** exploratory factor analysis lets the data reveal the number and structure of factors (new scale, no prior model); confirmatory factor analysis tests a pre-specified factor structure against new data (a form of structural equation modelling) and reports fit.
- **Principal component analysis:** a data-reduction method that repackages correlated variables into uncorrelated components ranked by variance explained; use it as a descriptive first pass or when maximising variance summarised is the goal; it differs from true factor analysis in seeking components that summarise, not latent factors that cause, the observed correlations.

## 13 · Diagnostic test statistics

Every screening or diagnostic test makes two kinds of promise and two kinds of mistake, and the whole apparatus of test statistics is a way of pinning numbers to those promises and mistakes. It is built on one small object: the **2x2 table** that cross-tabulates the true state of the world (disease present or absent, fixed by a reference or gold standard) against what the test said (positive or negative). The examiner reliably asks you to read that table, to compute an index from it, and above all to know which indices are intrinsic to the test and which shift with the population you apply it to (Kaplan & Sadock 2024, Bland 2015). Get that last distinction wrong and you will over-trust a positive screen in a low-risk clinic, the single most common error in this whole area.

The four cells carry standard labels (Kaplan & Sadock 2024). The test correctly flags a diseased person as a **true positive** (cell a) and correctly clears a healthy person as a **true negative** (cell d). It errs two ways: a **false positive** (cell b) labels a healthy person diseased, and a **false negative** (cell c) misses a diseased person. Read the table two ways: down the columns (given the true state, how did the test perform?) gives sensitivity and specificity; across the rows (given the test result,

what is the true state?) gives the predictive values. That change of direction is the heart of the topic.

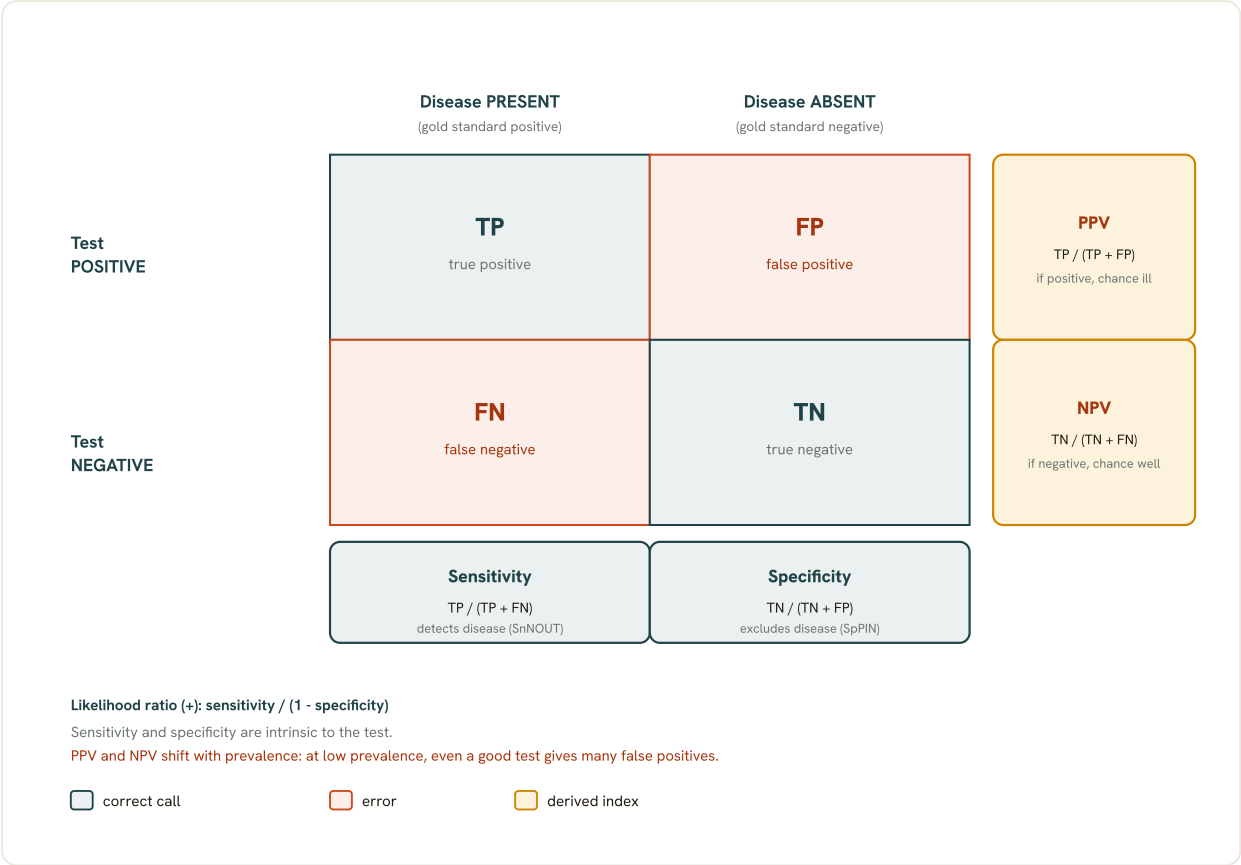



**Fig 7.5** The diagnostic 2x2 and its indices. Reading DOWN a column gives the intrinsic properties: sensitivity is the detection rate among the diseased, specificity the correct-negative rate among the well. Reading ACROSS a row gives the predictive values the clinician actually wants: the positive predictive value of a positive test and the negative predictive value of a negative one. The crucial caveat is that sensitivity and specificity stay fixed while the predictive values move with prevalence, which is why a positive screening test in a low-prevalence population is so often a false alarm.

## 13.1 The 2x2 table: the object everything is read from

**Fix the layout before you compute anything.** Disease status runs across the top (the gold-standard truth), test result down the side. Populate the four cells with absolute counts, never percentages, and always check that the marginal totals add up before you divide (Bland 2015). Every index below is just one ratio taken from this grid.

| TEST RESULT | DISEASE PRESENT | DISEASE ABSENT | ROW TOTAL |
|---|---|---|---|
| **Test positive** | True positive (a) | False positive (b) | **a + b** |
| **Test negative** | False negative (c) | True negative (d) | **c + d** |
| **Column total** | a + c (all diseased) | b + d (all non-diseased) | **N** |

Columns fix the true state and yield the intrinsic indices (down); rows fix the test result and yield the predictive values (across).

## 13.2 Sensitivity: detecting the diseased (SnNOUT)

**The test's ability to detect disease.** Sensitivity (use "sensitivity") is the proportion of truly diseased people the test correctly calls positive: true positives divided by all diseased (Kaplan &

Sadock 2024). It is read down the disease-present column, so it is a property of the test in diseased people and does not depend on how common the disease is.

#### Formula

Sensitivity = a / (a + c) = true positives / all diseased.

#### Complement

The false-negative rate = c / (a + c) = 1 − sensitivity: the fraction of diseased people the test misses.

#### Use

A sensitive test misses few cases, so it is what you want for a screening test or for ruling out a dangerous diagnosis.

**SnNOUT.** A highly **sensitive** test with a **Negative** result rules the disease **OUT**: if a test rarely misses disease, a negative from it makes disease unlikely. Hold the mnemonic SnNOUT (Sackett et al 2000).

RULE **SnNOUT rules out.** A negative result on a highly sensitive test rules the diagnosis out.

### 13.3 Specificity: clearing the healthy (SpPIN)

**The test's ability to exclude disease.** Specificity (use "specificity") is the proportion of truly non-diseased people the test correctly calls negative: true negatives divided by all non-diseased (Kaplan & Sadock 2024). It is read down the disease-absent column, so like sensitivity it is intrinsic to the test and independent of prevalence.

#### Formula

Specificity = d / (b + d) = true negatives / all non-diseased.

#### Complement

The false-positive rate = b / (b + d) = 1 − specificity: the fraction of healthy people wrongly flagged.

#### Use

A specific test rarely raises a false alarm, so it is what you want as a confirmatory test after a positive screen.

**SpPIN.** A highly **specific** test with a **Positive** result rules the disease **IN**: if a test rarely fires falsely, a positive from it makes disease likely. Hold the mnemonic SpPIN (Sackett et al 2000). The two indices trade off against each other along a threshold, which is exactly what an ROC curve plots (covered in the sampling and measurement notes).

### 13.4 Positive predictive value: trusting a positive result

**Of those who test positive, how many are truly diseased.** The positive predictive value (use "ppv", "positive predictive") reads across the test-positive row: of everyone the test flagged, what fraction actually has the disease (Kaplan & Sadock 2024). This is the number the clinician and the patient in front of you actually care about, and, crucially, it depends on prevalence.

#### Formula

PPV = a / (a + b) = true positives / all who test positive.

#### Reads

Across the top (test-positive) row, so it mixes the test's behaviour with how common the disease is in this population.

#### Moves with

Prevalence: the rarer the disease, the more the positive column is diluted by false positives, and the lower the PPV.

### 13.5 Negative predictive value: trusting a negative result

**Of those who test negative, how many are truly healthy.** The negative predictive value (use "npv", "negative predictive") reads across the test-negative row: of everyone the test cleared, what fraction really is disease-free (Kaplan & Sadock 2024). Like PPV it moves with prevalence, but in the opposite direction: in a low-prevalence setting almost everyone is truly healthy, so a negative result is nearly always right and NPV is high.

#### Formula

NPV = d / (c + d) = true negatives / all who test negative.

#### Reads

Across the bottom (test-negative) row.

#### Moves with

Prevalence, inversely: NPV rises as the disease becomes rarer.

### 13.6 The worked case: the same test, two prevalences

WORKED EXAMPLE

**Setup.** Take a test with sensitivity 90% and specificity 90%, and apply it first in a clinic where disease prevalence is 10%. Among 1000 people, 100 are diseased and 900 are not. Sensitivity 90% gives true positives a = 90 and false negatives c = 10. Specificity 90% gives true negatives d = 810 and false positives b = 90.

**Verify the intrinsic indices.** Sensitivity = a / (a + c) = 90 / 100 = 0.90; specificity = d / (b + d) = 810 / 900 = 0.90. Both reproduce the inputs, as they must.

**Predictive values.** PPV = a / (a + b) = 90 / (90 + 90) = 90 / 180 = 0.50, so a positive result carries only a 50% chance of true disease even with this good test. NPV = d / (c + d) = 810 / (810 + 10) = 810 / 820 = 0.988, about 99%.

**Now drop the prevalence to 1%.** Among 10000 people, 100 are diseased and 9900 are not. Sensitivity 90% still gives a = 90 and c = 10; specificity 90% now gives false positives b = 990 and true negatives d = 8910. Sensitivity and specificity are unchanged (90 / 100 and 8910 / 9900 = 0.90). But PPV = 90 / (90 + 990) = 90 / 1080 = 0.083, only about 8%. The very same test, in a low-prevalence population, turns nine out of ten positives into false alarms while its sensitivity and specificity never budge.

| 50% | 8% | 90/90 |
|---|---|---|
| PPV AT 10% PREVALENCE | PPV AT 1% PREVALENCE | SENS/SPEC, UNCHANGED |

### 13.7 Prevalence dependence: what is fixed and what moves

**The one distinction the examiner is really testing.** Sensitivity and specificity are **intrinsic** properties of the test: they are computed within the diseased and non-diseased columns respectively, so changing the mix of diseased to healthy people leaves them alone (Kaplan & Sadock 2024, Altman 1991). Predictive values are computed across the test-result rows, which are populated by both true and false cases, so they are hostages to **prevalence** (the pre-test probability). As prevalence falls, PPV falls (a good test still yields many false positives in a low-prevalence screen) and NPV rises; as prevalence climbs toward that of a specialist referral clinic, PPV rises and NPV falls. This is why a screening programme that looks superb in a high-risk clinic can generate a flood of false positives when rolled out to the general population (Wilson & Jungner 1968).

RULE **PPV rises with prevalence; sensitivity and specificity do not.** Predictive values move with the population; intrinsic indices are fixed to the test.

TRAP **A positive is not a confirmation.** A highly sensitive test with a positive result does not confirm disease in a low-prevalence population, where most positives are false; only a specific test's positive rules in.

### 13.8 The likelihood ratio and Fagan's nomogram

**The prevalence-proof way to update probability.** The likelihood ratio (use "likelihood ratio") folds sensitivity and specificity into a single number that says how much a given result shifts the odds of disease, and unlike the predictive values it is itself independent of prevalence (Sackett et al 2000, Bland 2015). The positive likelihood ratio, LR+ = sensitivity / (1 − specificity), is how much more likely a positive result is in a diseased than a non-diseased person; the negative, LR− = (1 − sensitivity) / specificity, does the same for a negative result.

**Pre-test to post-test.** The clinical value of an LR is that it converts a **pre-test** probability (use "pre-test", essentially the prevalence or your clinical suspicion before testing) into a **post-test** probability (use "post-test") after the result is known: post-test odds = pre-test odds x likelihood ratio. Fagan's nomogram (use Fagan for the eponym) draws this as three parallel scales; a ruled line from the pre-test probability through the likelihood ratio reads off the post-test probability without any arithmetic (Fagan 1975). For the worked test above, LR+ = 0.90 / (1 − 0.90) = 0.90 / 0.10 = 9, and LR− = (1 − 0.90) / 0.90 = 0.10 / 0.90 = 0.11. A rough reading: an LR+ above 10 or an LR− below 0.1 shifts probability decisively, while an LR near 1 barely moves it.

### 13.9 Hold the picture together

**Down the columns for the test, across the rows for the patient.** Sensitivity and specificity answer "how good is this test?" and are read down the true-state columns; predictive values answer "what does this result mean for this patient?" and are read across the result rows, dragging prevalence in with them; the likelihood ratio is the bridge that carries a pre-test probability to a post-test probability regardless of prevalence (Kaplan & Sadock 2024, Sackett et al 2000). If you can place a formula on each margin of the 2x2 and say in one breath which side prevalence touches, you have the topic.

**HOLD THIS**

Sensitivity and specificity are intrinsic to the test and do not move with prevalence; PPV and NPV do, so a positive result on even a good test means little in a low-prevalence population (SnNOUT rules out, SpPIN rules in).

**MNEMONIC**

### SnNOUT and SpPIN

A **Sn**sitive test, when **N**egative, rules the disease **OUT** (few misses, so a clear negative is trustworthy). A **Sp**ecific test, when **P**ositive, rules the disease **IN** (few false alarms, so a positive is trustworthy).

**PEARL**

Predictive values are what the patient hears, but they are borrowed from the population, not the test. Quote sensitivity and specificity to describe a test in the literature; compute PPV and NPV only for the specific prevalence you are working in; quote the likelihood ratio when you want a prevalence-proof measure to carry across settings.

**GOOD TO REMEMBER**

- **Sensitivity:** the proportion of diseased people the test calls positive; use it when you must not miss cases (screening, ruling out); sensitivity = a / (a + c) = true positives / all diseased; a Negative on a sensitive test rules OUT (SnNOUT); intrinsic, does not move with prevalence.
- **Specificity:** the proportion of healthy people the test calls negative; use it to confirm disease and avoid false alarms; specificity = d / (b + d) = true negatives / all non-diseased; a Positive on a specific test rules IN (SpPIN); intrinsic, does not move with prevalence.
- **Positive predictive value (PPV):** of those testing positive, the fraction truly diseased; use it to interpret a positive result in a given population; PPV = a / (a + b); rises with prevalence.
- **Negative predictive value (NPV):** of those testing negative, the fraction truly healthy; use it to interpret a negative result; NPV = d / (c + d); falls as prevalence rises.
- **Likelihood ratio (LR):** how much a result shifts the odds of disease, independent of prevalence; use it with Fagan's nomogram to move pre-test to post-test probability; LR+ = sensitivity / (1 − specificity), LR− = (1 − sensitivity) / specificity.
- **Prevalence (pre-test probability):** the proportion diseased before testing; recognise it as the lever that moves PPV and NPV while leaving sensitivity and specificity untouched; a low-prevalence screen yields many false positives even with a good test.

The prevalence figures used in a real screening decision come from population studies, so a test that performs well in a specialist clinic must be re-appraised at general-population prevalence before it is offered as a screen (Wilson & Jungner 1968); the prevalence of specific disorders is set out in the clinical chapters of Paper II, and the reliability and validity of a psychometric instrument itself, as distinct from these diagnostic indices, are covered in the Psychometry, Assessment and Descriptive Psychopathology notes.

## 14 · Statistical inference: p-values, confidence intervals, errors and power

Every significance test hands back two numbers, and the whole of applied statistical inference is the discipline of reading them without over-reading them. The first is the **p-value**, a verdict on chance; the second, ideally beside it, is the **confidence interval**, a statement of where the truth plausibly lies. The examiner is far less interested in whether you can compute either than in whether you know exactly what each does and does not mean, because the classic errors here are errors of interpretation, not arithmetic (Johnstone et al 2010, Bland 2015). Master four ideas and the topic is yours: what a p value actually measures, why the confidence interval is the more honest companion, the two ways an inference can be wrong, and the arithmetic of designing a study large enough to give a true effect a fair chance of being seen.

The **throughline** is the null hypothesis. A test assumes there is no difference (or no association) in the population, then asks how surprising the observed data would be under that assumption. The p value quantifies the surprise; the confidence interval turns the same information into a range of plausible effects. Errors and power then describe what can go wrong and how to guard against it before a single patient is recruited.

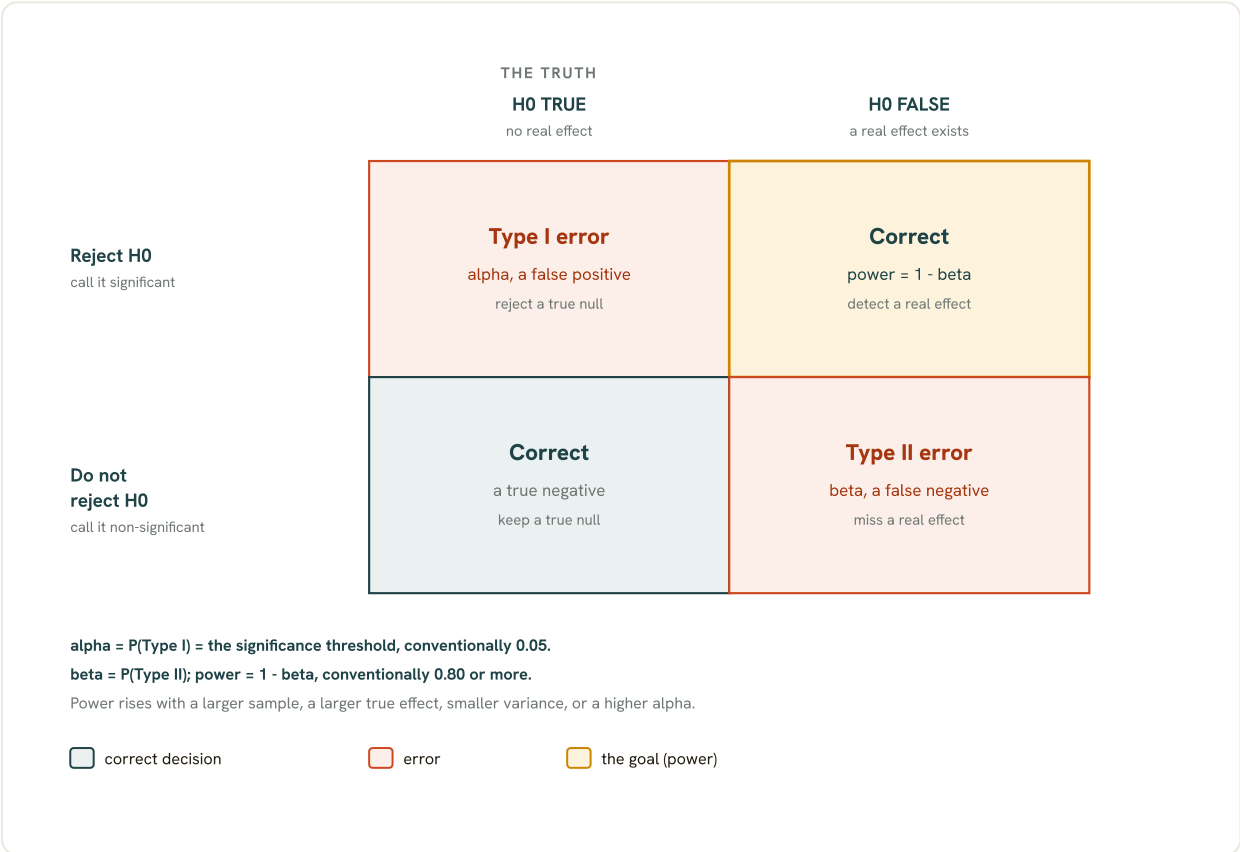



**Fig 7.6** Type I and type II error. A test can go wrong two ways. Rejecting a null that is actually true is a type I error, a false positive, and its probability is alpha, the significance level we set. Failing to reject a null that is actually false is a type II error, a false negative, and its probability is beta. Power, the ability to detect a real effect, is one minus beta, and an underpowered study is the commonest reason a true effect is missed.

### 14.1 The p-value: what it is

**A probability about the data, given the null.** The p-value (also written p value) is the probability of obtaining the observed result, or one more extreme, if the null hypothesis of no true difference

were correct (Johnstone et al 2010, Altman 1991). It is a measure of how compatible the data are with the null: a small p value says the data would be a rare fluke under the null, so we doubt the null. It is emphatically a probability about the data assuming the null is true, not a probability about the null itself.

### 14.2 The alpha threshold and significance

**Where the line is drawn.** Before the study, an investigator fixes a significance threshold, alpha, the p value at or below which the null will be rejected. By convention alpha = **0.05** (Johnstone et al 2010, Bland 2015). A result with p at or below 0.05 is called statistically significant: under a true null we would see a result this extreme less than 1 time in 20, which we treat as evidence against the null. The 0.05 line is a convention, not a law of nature, and stricter thresholds (0.01, or a Bonferroni-adjusted value across many tests) are used where false positives are costly. Alpha is chosen by the investigator; the p value is computed from the data.

RULE **Set alpha first, then read p.** Alpha is your pre-declared tolerance for a false positive; the p value is what the data return. Compare the one to the other, never move the goalposts after seeing the data.

### 14.3 What a p-value is NOT

**The three misreadings the examiner hunts for.** A p value is not the probability that the null hypothesis is true, and it is not the probability that the result arose by chance in any absolute sense; it is conditional on the null being true (Altman 1991, Sterne & Davey Smith 2001). Nor does a p value measure the size or the importance of an effect: a vanishingly small p value can attach to a trivial difference if the sample is huge, and a non-significant p can hide a large effect measured imprecisely. And a non-significant result (p above 0.05) is not proof that the null is true, absence of evidence is not evidence of absence. Hold these three: p is not the probability the null is true, p is not the effect size, and a large p does not prove no effect.

TRAP **p is not P(null is true).** A p of 0.03 does not mean a 3% chance the null holds; it means data this extreme would occur 3% of the time if the null held. The direction of the conditional is the whole trap.

### 14.4 The confidence interval: the plausible range

**From a yes/no verdict to a range of effects.** A confidence interval is the range of plausible values within which the true population value is likely to lie, given the data (Johnstone et al 2010, Altman 1991). A **95 percent CI** is the interval such that, over many repeated samples, 95% of the intervals so constructed would contain the true value; loosely, we can be 95 percent confident the true effect lies inside it. Where the p value gives a bare significant/non-significant verdict, the CI shows both the best estimate (the point estimate at its centre) and its precision (its width). A trial that finds chlorpromazine beats placebo with a relative risk of 2.3 (95% CI 1.2 to 3.5) tells you not only that the effect is real but that the true effect plausibly ranges from modest to large (Johnstone et al 2010).

### 14.5 Reading significance off a confidence interval

**The no-effect value inside the interval.** A CI carries the significance test within it, read against the value of no effect. For a **difference** of means (or any subtracted quantity) the no-effect value is **0**: if the 95% CI for the difference crosses 0, the result is non-significant at the 0.05 level. For a **ratio** measure, a relative risk (RR) or odds ratio (OR), the no-effect value is **1** (equal risk, equal odds): if the 95% CI for the ratio crosses 1, the result is non-significant (Johnstone et al 2010, Altman 1991). So an RR of 2.3 with CI 1.2 to 3.5 is significant (the interval sits wholly above 1), whereas an OR of 1.4 with CI 0.9 to 2.1 is not (the interval straddles 1). The CI is more informative than the p value alone because it delivers the significance verdict and the effect size and its precision in a single line.

RULE **Cross 0 for a difference, cross 1 for a ratio.** A 95% CI for a difference that includes 0, or for an RR/OR that includes 1, means the result is not statistically significant.

### 14.6 Type I and Type II error

**The two ways an inference can be wrong.** Because inference is a decision under uncertainty, two mistakes are possible against the two states of the truth (Johnstone et al 2010, Bland 2015).

#### Type I error

A false positive: rejecting a null hypothesis that is actually true, that is, declaring a difference that does not exist. Its probability is alpha, so setting alpha at 0.05 caps the false-positive rate at 5%.

#### Type II error

A false negative: failing to reject a null hypothesis that is actually false, that is, missing a difference that truly exists. Its probability is termed beta.

**The 2x2 of truth against decision.** Cross the true state of the world (null true, or null false) with the decision taken (do not reject, or reject) and four cells result: two correct (a true negative, and a true positive), and the two errors above. The type I error sits where a true null is rejected; the type II error where a false null is retained. Tightening alpha to reduce type I error, all else equal, raises beta and so raises type II error: the two trade off against each other.

### 14.7 Power and its four drivers

**The chance of finding a real effect.** The power of a test is 1 minus beta, the probability that it will detect a true effect of a given size when one genuinely exists (Tasman et al 2015, Johnstone et al 2010). Colloquially, power is the chance of finding what you are looking for. The convention is to design for power of **80** percent or more, meaning at most a 20% chance of missing a true effect (beta = 0.2). Power is fixed by four interlocking quantities, and moving any one moves power:

#### Alpha

The significance threshold. A stricter (smaller) alpha lowers power, all else equal.

#### Sample size

The number of participants. A larger sample sharpens the estimate and raises power; too small a sample is the commonest reason a true effect is missed.

### Effect size

The magnitude of the true difference sought. Larger effects are easier to detect, so they need smaller samples.

### Variance

The scatter of the data. Less variability sharpens the estimate and raises power.

Fix any four of alpha, power, effect size, variance and sample size and the fifth is determined, which is exactly how a sample-size calculation works.

TRAP **A non-significant result may be an underpowered study.** Failure to reject the null can mean there is no effect, or that the study was too small to see one. The two are not the same, and the p value alone cannot tell them apart.

## 14.8 Effect size and Cohen's d

**Standardising the size of a difference.** An effect size expresses the magnitude of a difference in units that are comparable across studies and scales, independent of sample size (New Oxford Textbook 2020). For a difference between two means the standard measure is cohen's d, the difference in means divided by the pooled standard deviation. Cohen's conventional benchmarks are small = **0.2**, medium = **0.5**, and large = **0.8** (New Oxford Textbook 2020). For a correlation, the effect size is r itself (or r-squared, the shared variance). Effect size is the input that most strongly governs the sample a study needs: a large expected effect can be caught with few participants, a small one demands many.

## 14.9 Sample size, the MCID and the power calculation

**Sizing the study before it runs.** A sample size calculation is performed at the design stage to find the number of participants needed to give a study adequate power (Johnstone et al 2010, Altman 1991). Its inputs are the other four quantities: the chosen alpha (usually 0.05), the target power (usually 0.8), the smallest effect worth detecting, and an estimate of the variance or event rate (from pilot data or the literature). Feed these in and the power calculation returns the required n; equivalently, given a fixed n, it returns the power to detect a stated effect. The effect fed in should be the minimum clinically important difference, the smallest change that would matter to a patient or alter management, not merely the smallest that could be made statistically significant. Powering a trial for a statistically detectable but clinically trivial difference wastes participants; powering it for the minimum clinically important difference sizes it to answer a question worth asking.

| 0.05 | 80% | 0.2 / 0.5 / 0.8 |
|---|---|---|
| ALPHA | POWER | COHEN D SMALL/MED/LARGE |

## 14.10 A worked read: p, CI and the four cells together

WORKED EXAMPLE

**Reading a trial result.** A relapse-prevention trial reports a relative risk of relapse on active drug versus placebo of RR = 2.3, 95% CI 1.2 to 3.5, p = 0.01 (Johnstone et al 2010). Read it in three moves. First the p value: under a true null of no difference, a result this extreme would occur about 1 time in 100, below alpha 0.05, so we reject the null and call it significant. Second the confidence interval: the 95% CI runs 1.2 to 3.5 and does not cross the ratio no-effect value of 1, which confirms significance, and its width tells us the true effect plausibly ranges from modestly (1.2) to substantially (3.5) raised, an imprecise estimate a larger sample would tighten. Third the error frame: in rejecting the null we risk a type I error, capped at alpha = 0.05; had the CI instead been 0.8 to 4.1 (crossing 1, non-significant) we could not conclude no effect, because a small, underpowered study courts a type II error of missing a real difference.

**A difference, not a ratio.** Had the outcome been a mean symptom-score difference of 2.0 points with 95% CI −0.5 to 4.5, the no-effect value is 0, the interval crosses it, and the result is non-significant, whatever the point estimate suggests.

COMMON TRAP

Statistical significance is not clinical importance. A trial of 10,000 patients can return $p < 0.001$ for a 1-point difference on a 60-point depression scale, a difference no patient would notice; conversely a genuinely useful effect in a small study may miss significance purely for want of power. Always read the effect size and its confidence interval, not the p value alone, and anchor the judgement to the minimum clinically important difference.

MNEMONIC

**Type I is the FALSE ALARM (reject a true null); Type II is the MISS (retain a false null).**

A smoke alarm going off with no fire is a type I error (false positive, probability alpha); a fire the alarm sleeps through is a type II error (false negative, probability beta). Power (1 minus beta) is the alarm actually sounding for a real fire.

HOLD THIS

A 95% CI that crosses 0 (for a difference) or 1 (for a ratio) is non-significant; type I error = alpha = 0.05 (false positive), type II error = beta (false negative), power = 1 minus beta, conventionally 80% or more.

**GOOD TO REMEMBER**

- **p-value:** the probability of the observed data or more extreme if the null were true; use it to judge compatibility with the null; the significance threshold is alpha = 0.05.
- **Confidence interval (95% CI):** the plausible range for the true population value; use it instead of the p value alone because it shows effect size and precision; non-significant if it crosses 0 (difference) or 1 (ratio RR/OR).
- **Type I error:** a false positive, rejecting a true null; guarded by the significance threshold; its probability is alpha (conventionally 0.05).
- **Type II error:** a false negative, retaining a false null; arises chiefly from too small a sample; its probability is beta.
- **Power:** the ability to detect a true effect; design for it before recruiting; power = 1 minus beta, conventionally 80% or more.
- **Effect size (Cohen's d):** the standardised magnitude of a difference; use it to compare across studies and drive sample size; d = 0.2 small, 0.5 medium, 0.8 large.
- **Sample size / power calculation:** the design-stage arithmetic sizing a study; use it to secure adequate power; its inputs are alpha, power, the minimum clinically important effect size, and the variance.

For the derivations and worked confidence-interval formulae see Altman (1991) and Bland (2015); the interpretation traps are set out plainly in Sterne & Davey Smith (2001). Where these inference tools drive the choice of a specific significance test, that selection is treated in the sibling tests-of-significance material; twin, family and adoption designs that generate the effect estimates fed into these calculations are covered in the Psychiatric Genetics notes, and the reliability and validity of the measuring instrument itself (as distinct from a study's statistical validity) belong to the Psychometry, Assessment and Descriptive Psychopathology notes.

## 15 · Survival analysis, NNT and ROC

Three tools cluster here because each squeezes a clinically useful single number out of a slightly awkward kind of data. Survival analysis handles data where the outcome is **time to an event** and some patients have not had the event when the study ends. The number needed to treat converts a trial's risk difference into a bedside count of patients. The ROC curve measures how well a diagnostic test separates the sick from the well across every possible cut-off. The examiner tests interpretation, not derivation: what the curve, the count and the area under the curve mean, and the one trap that goes with each (Kaplan & Sadock 2024, Bland 2015).

**Why time-to-event needs its own method.** If you simply compared the proportion relapsed at two years between two arms, you would throw away everyone who left the study early or who had not relapsed at the cut-off. Survival analysis keeps them by treating their incomplete follow-up as partial information rather than as a missing value to be discarded (Kaplan & Sadock 2024).

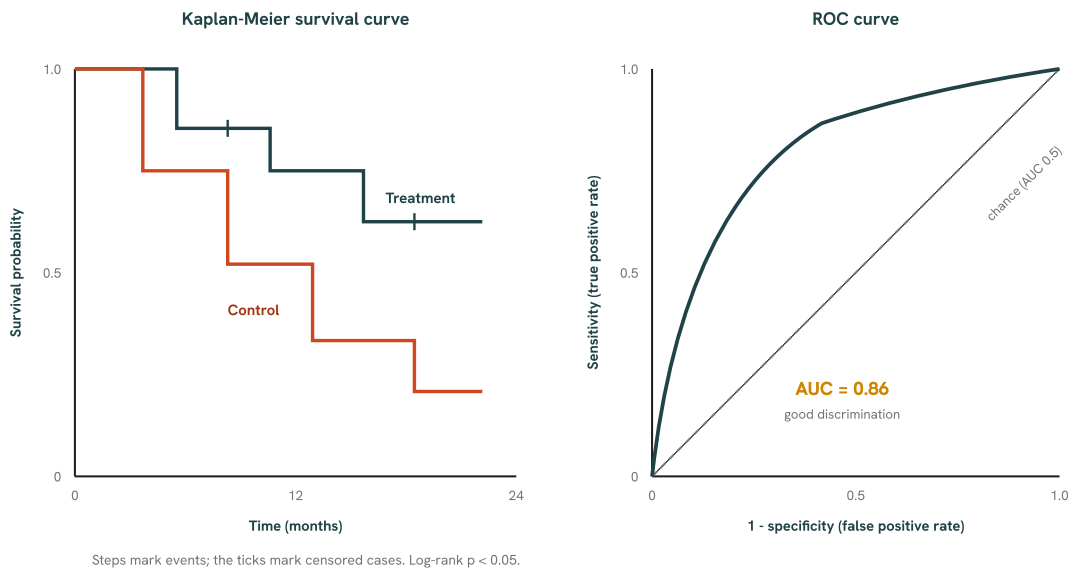



**Fig 7.7** Survival and ROC curves. The Kaplan-Meier curve plots the probability of remaining event-free over time, stepping down at each event, with ticks for censored cases; two curves are compared with the log-rank test and summarised by the hazard ratio. The ROC curve plots sensitivity against one minus specificity across every cut-off of a test: the further it bows toward the top-left corner the better, and the area under it (AUC) grades overall discrimination from 0.5 (chance) to 1.0 (perfect).

## 15.1 Time-to-event data and censoring

**What survival data are.** Many psychiatric questions live in the time domain: time to onset of drug effect, time in hospital, time to remission or to relapse (Kaplan & Sadock 2024). These are generically **time-to-event** or survival variables; the original event studied was time to death, which gives the field its name. The two things recorded for each subject are the length of follow-up and whether the event had occurred by the end of it.

**Censoring.** A subject is censoringed (a censored observation) when the event has not occurred by the time observation stops: the study ends, or the patient is lost to follow-up, or withdraws (Kaplan & Sadock 2024). We know only that their true time-to-event exceeds the time we watched them. This is right-censoring, the common form. Crucially, a censored patient is not the same as a dropout to be deleted: they contribute information for the whole period they were observed, and are only removed from the risk set after the point at which we lose sight of them.

TRAP **Censoring is not ignored loss to follow-up.** Censoring is not the same as loss to follow-up being ignored: Kaplan-Meier accounts for it by using each censored case up to the moment it is censored, then dropping it from the denominator, rather than discarding it or coding it as an event.

## 15.2 The Kaplan-Meier survival curve

**The step curve everyone draws.** The kaplan-meier estimator is the nonparametric way to draw the survival function from the data: it makes no assumption about the shape of the underlying distribution (Kaplan & Sadock 2024). It splits time at each point where an event actually occurs and, at each such point, multiplies the running survival probability by the proportion of those

still at risk who survived that instant. The probability of surviving beyond time t is therefore the product of the conditional survival probabilities across all preceding intervals, which is why it is a step function that drops only when an event happens, not at fixed calendar intervals.

**Reading the curve.** The y-axis is the cumulative proportion still event-free; the curve starts at 1.0 and steps down. A small vertical tick on the curve marks a **censored** case (a patient still event-free when last seen). The height where the curve crosses 0.5 gives the **median survival time**, the single summary statistic most often quoted (Kaplan & Sadock 2024). Flatter curve means better survival.

### 15.3 Comparing two curves: the log-rank test

**Is the gap between two curves real?** Two survival curves, for example two treatment arms, are compared with the log-rank test, the most commonly used nonparametric comparison (Kaplan & Sadock 2024, Bland 2015). At every event time it compares the number of events observed in each group with the number expected if the two curves shared one underlying survival experience, then sums those observed-minus-expected contributions across all event times into a single chi-square-type statistic. Its null hypothesis is that the two survival distributions are identical. It uses the whole curve, not survival at one arbitrary time-point, and weights every event time equally (the related Gehan and generalized Wilcoxon tests weight early events more).

RULE **Curves to log-rank, effect size to Cox.** Use the log-rank test to ask whether two Kaplan-Meier curves differ; use Cox regression to put a number, the hazard ratio, on how much they differ and to adjust for covariates.

### 15.4 The hazard ratio from Cox regression

**The instantaneous relative risk.** The hazard rate is the probability of the event happening at time t given the subject has survived event-free up to t: an instantaneous risk. Cox proportional hazards regression models this time-to-event outcome against covariates without assuming a distributional shape (it is semiparametric), and yields the hazard ratio (Kaplan & Sadock 2024). The hazard ratio is the ratio of the hazard in one group to the hazard in the other: read it like a relative risk that applies at every instant. A hazard ratio of **1.0** means no difference in the rate of events; above 1 the exposure or treatment raises the instantaneous event rate, below 1 it lowers it. The name comes from the **proportionality assumption**: the model assumes the two hazards stay in constant proportion over time (Kaplan & Sadock 2024).

### 15.5 The number needed to treat

**From a risk difference to a headcount.** The number needed to treat (nnt) translates a trial result into a clinical unit a resident can quote at the bedside: how many patients must receive the treatment for one extra good outcome, compared with the control (Cooper & Jacoby 2013, Bland 2015). It is the reciprocal of the **absolute risk reduction** (ARR), the difference between the control event rate and the experimental event rate:

#### Absolute risk reduction (ARR)

Control event rate minus experimental event rate, expressed as a proportion (a difference in the probability of the bad outcome).

### Number needed to treat (NNT)

NNT = 1 / ARR, rounded up to the next whole number; the count of patients to treat to prevent one extra bad outcome.

### Number needed to harm (NNH)

The mirror image: 1 / absolute risk increase in an adverse outcome; the count of patients to treat before one extra harm occurs.

**Direction of merit.** A large NNT means many patients must be treated for a single benefit, so the treatment is weak; a small NNT means the benefit comes cheaply, so the treatment is strong. As a rough guide an NNT under **10** denotes a clinically useful treatment effect, and NNTs for psychiatric interventions are broadly comparable to those in general medicine (Cooper & Jacoby 2013). NNH is best large (harm is rare); NNT is best small.

RULE **Smaller NNT, stronger treatment.** A smaller NNT is a stronger treatment: fewer patients treated per good outcome gained. A small NNT and a large NNH together is the ideal profile.

## 15.6 A worked NNT

WORKED EXAMPLE

**CBT for residual depression (Paykel et al 1999, via Cooper & Jacoby 2013).** Patients with residual depressive symptoms were randomised to clinical management alone or clinical management plus cognitive behaviour therapy. Over follow-up the relapse rate was **47%** in the clinical-management (control) arm and **29%** in the CBT arm. The absolute risk reduction is the difference in these relapse rates: ARR = 47% − 29% = **18%** (0.18). The NNT is the reciprocal: NNT = 1 / 0.18 = 5.56, rounded up to **6**. Reading: about six patients with residual depressive symptoms must be treated with CBT to prevent one extra relapse. Since 6 is under 10, this is a clinically worthwhile effect.

## 15.7 The ROC curve

**A test across all its thresholds.** A continuous diagnostic test (a rating-scale score, a biomarker level) gives a different sensitivity and specificity at every cut-off you might pick. The roc (receiver operating characteristic) curve plots this whole family of trade-offs on one graph: the **sensitivity** (the true positive rate) on the y-axis against **1 minus specificity** (the false positive rate) on the x-axis, traced across all thresholds (Altman 1991, Bland 2015). Lowering the cut-off catches more true cases (sensitivity rises) at the cost of more false alarms (specificity falls), so the curve sweeps from the bottom-left corner up to the top-right. A test with no discriminating power sits on the diagonal from (0,0) to (1,1); a good test bows up towards the top-left corner, where sensitivity is high and false positives are few.

## 15.8 The area under the curve (AUC)

**One number for overall discrimination.** The auc, the area under the ROC curve, summarises the whole curve as a single index of discrimination: the probability that the test will score a randomly chosen affected patient higher than a randomly chosen unaffected one (Altman 1991). Its scale is anchored at two points a resident must hold: an AUC of **0.5** is the diagonal, chance-level

discrimination (no better than a coin toss), and an AUC of **1.0** is a perfect test that separates the two groups completely. In between, higher is better, so AUC lets you rank competing tests on one axis.

| 0.5 | 1.0 | < 10 |
|---|---|---|
| AUC AT CHANCE | AUC PERFECT TEST | NNT: USEFUL EFFECT |

### 15.9 Choosing a cut-off from the curve

**Where to set the threshold.** The ROC curve also guides the practical choice of a single operating cut-off (Altman 1991, Bland 2015). The point closest to the top-left corner gives the best overall balance of sensitivity and specificity, but the clinically right cut-off depends on the costs of the two error types. For a screening test, where missing a case is worse than a false alarm, slide the cut-off to favour **sensitivity** (accept more false positives); for a confirmatory test, where a false positive is costly, favour **specificity**. The curve does not choose for you; it shows the trade-off you are buying at each threshold. The reliability and validity of the underlying instrument itself are covered in the Psychometry, Assessment and Descriptive Psychopathology notes.

TRAP **The diagonal is not a bad test drawn badly.** An ROC curve sitting on the 45-degree line (AUC 0.5) is not an error of plotting: it is a test with zero discriminating ability. Only a curve bowing above the diagonal carries diagnostic information.

HOLD THIS

Kaplan-Meier steps down and keeps censored cases (median survival where it hits 0.5); log-rank compares two curves, Cox gives the hazard ratio (1.0 = no effect); NNT = 1 / ARR and smaller is stronger; ROC plots sensitivity against 1 minus specificity, and AUC runs 0.5 (chance) to 1.0 (perfect).

**GOOD TO REMEMBER**

- **Survival analysis:** the family of methods for time-to-event data with censoring; use when the outcome is time until an event and some subjects have not had it by study end; the key summary is the median survival time.
- **Censoring:** incomplete follow-up where the event has not yet occurred; use it (do not delete it) whenever a subject is lost, withdraws, or the study ends event-free; it contributes to the risk set only up to the point of censoring.
- **Kaplan-Meier curve:** the nonparametric step estimate of the survival function; use to picture and summarise survival in one or more groups; it drops only at observed event times and marks censored cases with ticks.
- **Log-rank test:** the nonparametric test comparing two or more survival curves; use to ask whether the curves differ overall; null hypothesis is identical survival distributions, and it weights all event times equally.
- **Hazard ratio (Cox regression):** the ratio of instantaneous event rates between groups from a semiparametric model; use to quantify and covariate-adjust the difference between curves; a hazard ratio of 1.0 means no difference.
- **Number needed to treat (NNT):** the count of patients treated per one extra good outcome; use to translate a trial's ARR into clinical terms; NNT = 1 / absolute risk reduction, and under 10 is a useful effect.
- **Number needed to harm (NNH):** the count of patients treated before one extra adverse event; use to size the harm of a treatment; NNH = 1 / absolute risk increase, and larger is safer.
- **ROC curve:** the plot of sensitivity against 1 minus specificity across all thresholds; use to show a continuous test's sensitivity-specificity trade-off and to pick a cut-off; the best-balance point sits nearest the top-left corner.
- **AUC (area under the ROC curve):** a single index of overall discrimination; use to rank or compare diagnostic tests; it runs from 0.5 (chance) to 1.0 (perfect).

For survival methods and Cox regression see Kaplan & Sadock (2024); for ARR, NNT and NNH see Cooper & Jacoby (2013); for ROC and AUC see Altman (1991) and Bland (2015). Genetic time-to-onset study designs are treated in the Psychiatric Genetics notes, and the disorder-specific efficacy figures these tools summarise appear in the clinical chapters of Paper II.

## Epidemiology

### 16 · Measures of disease frequency and association

Epidemiology begins with two questions and answers each with a number: how common is a disorder, and what raises the chance of getting it. The first question is answered by the **measures of disease frequency**, of which the two workhorses are incidence and prevalence; the second by the **measures of association**, which quantify how much an exposure changes the risk of an outcome (Kaplan & Sadock 2024). The examiner reliably asks you to separate a count of new cases from a count of existing cases, to know which measure of association comes from which study design, and to read an association off a 2x2 table without confusing risk with odds. Get the incidence-versus-prevalence distinction wrong and every downstream inference about causation and burden goes with it.

This material owns the measures themselves and the national numbers; the **epidemiology of specific disorders** (the prevalence figures for schizophrenia, mood disorders, and the rest) is developed in the clinical chapters of Paper II. Here we keep the tools, verified formula by formula, and the study design each tool is native to.

## 16.1 Prevalence: the stock of existing cases

**Existing cases in a population at a point or over a period.** Prevalence (use "prevalence") is a ratio: the number of people who have the disorder divided by the number in the population, over a designated time window (Kaplan & Sadock 2024). It is a snapshot of the burden of disease, so it counts both old and new cases indiscriminately and depends heavily on how long the disorder lasts. Three elements fix any prevalence estimate: the case definition, the target population, and the time.

### Point prevalence

Point prevalence (use "point prevalence"): the proportion of a population who have the disorder at a single specified point in time.

### Period prevalence

Period prevalence (use "period prevalence"): the proportion who have the disorder at any time within a stated interval (1-month, 1-year, or lifetime prevalence).

### Lifetime prevalence

The proportion who have had the disorder at any point in their lives; useful for risk-factor studies but subject to recall bias.

For mental disorders, point prevalence is hard to pin down because symptoms wax and wane daily, so the **1-month** and **1-year** period prevalences are used in practice; the 1-year figure is what service planners multiply by the population to forecast caseload (Kaplan & Sadock 2024).

## 16.2 Incidence: the flow of new cases

**New cases arising over a period, per population at risk.** Incidence (use "incidence") counts only fresh cases: its numerator is the number of new events arising in a specified time period, and its denominator is the population at risk of having that event (Kaplan & Sadock 2024). People who already have the disorder at the start are removed from the denominator because they are, by definition, no longer at risk of developing it. Incidence is the measure of choice for a first-onset question, for example the number of new cases of post-traumatic stress disorder after a natural disaster.

### Cumulative incidence

Cumulative incidence (use "cumulative incidence"): new cases divided by the number of disease-free people at risk at the start of the period; a proportion (risk), so it needs a stated time period to mean anything.

### Incidence rate

Incidence rate (use "incidence rate"): new cases divided by the total person-time at risk; a true rate with units of per person-year, robust to people entering and leaving the population.

### Person-time

The sum, over all individuals, of the time each was actually at risk until they became a case or were lost to follow-up; the denominator of the incidence rate.

Because most mental disorders are rare per year, estimating incidence needs large populations at risk followed for long periods, and most incidence rates are quoted over a 1-year window (Kaplan & Sadock 2024).

RULE **Incidence is a rate of the new, prevalence is a stock of the existing.** Ask "new cases over a period" for incidence, "all cases at a moment" for prevalence.

### 16.3 The bridge: prevalence is roughly incidence times duration

**The equation that ties the two frequency measures together.** Prevalence, incidence, and mean duration are mathematically related: the prevalence of a disorder is proportional to its incidence multiplied by its mean duration, written **prevalence is roughly incidence times duration** (P is proportional to I x d) for a stable, low-prevalence condition (Kaplan & Sadock 2024). The consequence is the single most examined trap in this topic. Prevalence reflects not only how many new cases appear but also how long cases stay ill, so anything that lengthens duration inflates prevalence even when incidence is flat.

TRAP **Prevalence rises with longer illness duration even if incidence is unchanged.** A treatment that prolongs life without curing (the AIDS example) raises prevalence; a cure lowers it. Rising prevalence never by itself proves rising incidence.

COMMON TRAP

Do not read a climbing prevalence as more people falling ill. Because prevalence is roughly incidence times duration, better survival or chronification lifts prevalence with incidence untouched. When you need to tell a genuine rise in new cases apart from improved survival, you must use an incidence measure, not prevalence.

### 16.4 From frequency to association: the 2x2 table

**Every measure of association is one ratio read off this grid.** The **measures of association** compare disease frequency between an exposed and an unexposed group, and all of them are computed from a single 2x2 table that cross-tabulates exposure against outcome (Kaplan & Sadock 2024). Fix the layout first, populate the four cells with absolute counts, and check the marginal totals before you divide. In the standard orientation the exposed row holds cells a and b, the unexposed row cells c and d.

| GROUP | DISEASE PRESENT | DISEASE ABSENT | ROW TOTAL |
|---|---|---|---|
| **Exposed** | a | b | **a + b** |
| **Unexposed** | c | d | **c + d** |

Risk in the exposed reads across the top row, a / (a + b); risk in the unexposed reads across the bottom row, c / (c + d).

### 16.5 Relative risk: the ratio of two incidences, from a cohort

**How many times more likely the exposed are to develop disease.** The relative risk (use "relative risk") is the ratio of two incidences: the incidence in the exposed divided by the incidence in the unexposed (Kaplan & Sadock 2024). Because it needs the actual risk (incidence) in each group, it can only be computed when you have followed disease-free people forward and watched new cases arise, that is, from a cohort study. A relative risk greater than 1 marks the exposure as a risk factor; less than 1, a protective factor; exactly 1, no association.

#### Formula

Relative risk = incidence in exposed / incidence in unexposed = [a / (a + b)] / [c / (c + d)].

#### Design

Cohort study (follows exposed and unexposed forward), because only a cohort yields incidence directly.

#### Reads

A risk ratio: "the exposed are RR times as likely to develop the disorder as the unexposed."

---

RULE **Relative risk comes from a cohort, the odds ratio from a case-control.** If the design gives you incidence you can quote a relative risk; if it starts from cases and controls you are limited to an odds ratio.

---

### 16.6 Odds ratio: the case-control measure that approximates RR when disease is rare

**The ratio of the odds of exposure, computed when you cannot measure incidence.** The odds ratio (use "odds ratio") is used when the study begins with cases and controls rather than exposed and unexposed cohorts, so incidence and therefore relative risk cannot be measured directly (Kaplan & Sadock 2024). It compares the odds of disease in the exposed (a / b) with the odds of disease in the unexposed (c / d), which reduces to the cross-product ad / bc. Like the relative risk it varies around 1: above 1 an association with disease, below 1 an association with not having disease, exactly 1 no association.

#### Formula

Odds ratio = (a / b) / (c / d) = ad / bc (the cross-product ratio).

#### Design

Case-control study (starts from diseased cases and disease-free controls, then looks back at exposure).

#### Key property

The odds ratio approximates the relative risk when the outcome is rare (the rare-disease assumption), so for uncommon disorders a case-control odds ratio stands in for a cohort relative risk.

## 16.7 Worked read of a 2x2: relative risk and odds ratio side by side

WORKED EXAMPLE

**Setup.** A cohort of 200 exposed and 200 unexposed people is followed. Among the exposed, 40 develop the disorder and 160 do not, so a = 40, b = 160. Among the unexposed, 10 develop it and 190 do not, so c = 10, d = 190.

**Relative risk.** Incidence in exposed = a / (a + b) = 40 / 200 = 0.20. Incidence in unexposed = c / (c + d) = 10 / 200 = 0.05. Relative risk = 0.20 / 0.05 = **4.0**, so the exposed are four times as likely to develop the disorder.

**Odds ratio.** Odds of disease in exposed = a / b = 40 / 160 = 0.25. Odds in unexposed = c / d = 10 / 190 = 0.0526. Odds ratio = ad / bc = (40 x 190) / (160 x 10) = 7600 / 1600 = **4.75**.

**The lesson.** With a disease this common (20 percent in the exposed) the odds ratio (4.75) overstates the relative risk (4.0): the odds ratio always lies further from 1 than the risk ratio. Make the disease rare (say 5 new cases in the exposed rather than 40) and the two converge, which is exactly the rare-disease approximation in action.

| 4.0 | 4.75 | 1 |
|---|---|---|
| RELATIVE RISK (COHORT) | ODDS RATIO (CROSS-PRODUCT) | VALUE MEANING NO ASSOCIATION |

## 16.8 Attributable risk and the population attributable risk

**The excess risk, not the ratio of risks.** Where relative risk and odds ratio are ratios, the attributable risk (use "attributable risk") is a difference: the excess risk of disease in the exposed that can be attributed to the exposure, calculated as the incidence in the exposed minus the incidence in the unexposed (Kaplan & Sadock 2024). It answers "how much extra disease does this exposure add", and expressed as a fraction of the exposed incidence it becomes the attributable fraction, the share of disease in exposed people that the risk factor explains.

### Attributable risk

Excess risk in the exposed = incidence in exposed minus incidence in unexposed.

### Population attributable risk

The excess disease in the whole population attributable to the exposure; it depends on how common the exposure is, so a rare exposure yields a small population attributable risk however strong its relative risk.

### Public-health use

If the exposure is truly causal, the population attributable risk estimates the disease reduction achievable by removing the exposure.

In the worked cohort, attributable risk = 0.20 minus 0.05 = **0.15**, meaning 15 excess cases per 100 exposed persons are attributable to the exposure; the ratio measures (RR, OR) tell you the strength of association, the attributable measures tell you the burden.

## 16.9 The measures assembled: formula, design, interpretation

**One table to hold the whole toolkit.** The examiner wants you to match each measure to its formula, the study design it is native to, and how to read it (Kaplan & Sadock 2024, Bland 2015).

| MEASURE | FORMULA | DESIGN IT COMES FROM | INTERPRETATION |
|---|---|---|---|
| **Point prevalence** | cases at a point / population | **Cross-sectional survey** | Burden now; stock of existing cases |
| **Period prevalence** | cases in interval / population | **Cross-sectional / repeated survey** | Burden over an interval (1-month, 1-year, lifetime) |
| **Cumulative incidence** | new cases / at-risk at start | **Cohort** | Risk of new disease over the period |
| **Incidence rate** | new cases / person-time at risk | **Cohort** | Speed of new-case occurrence, per person-year |
| **Relative risk** | incidence exposed / incidence unexposed | **Cohort** | How many times more likely disease is in exposed; RR > 1 harmful |
| **Odds ratio** | ad / bc | **Case-control** | Approximates RR when disease is rare; OR > 1 associated with disease |
| **Attributable risk** | incidence exposed minus incidence unexposed | **Cohort** | Excess (extra) risk in the exposed due to exposure |
| **Population attributable risk** | excess disease in population from exposure | **Cohort / population data** | Disease preventable by removing exposure; depends on exposure prevalence |

Ratio measures (RR, OR) give strength of association; difference measures (attributable risk) give the excess burden; frequency measures (prevalence, incidence) give how common the disorder is.

### 16.10 The national numbers this material owns

**Where the population figures come from.** The flagship Indian frequency estimate is the National Mental Health Survey of India (2015 to 2016), a cross-sectional survey that reported a current weighted prevalence of any mental disorder of about **10.6** percent and a lifetime prevalence near **13.7** percent, with a treatment gap for mental disorders exceeding **70** percent (Gururaj et al 2016). These are prevalence numbers, a stock of existing cases at survey time, not incidence; they belong here as measures of frequency, while the disorder-by-disorder breakdown is carried in the clinical chapters of Paper II.

**INDIA**

The National Mental Health Survey of India (2015 to 2016) gives a current weighted prevalence of any mental disorder of about **10.6** percent (lifetime near **13.7** percent) and a treatment gap above **70** percent. Read these as point/period prevalence figures: they measure how many people have a disorder now, and because prevalence is roughly incidence times duration they cannot on their own tell you whether new cases are rising.

**HOLD THIS**

Incidence counts new cases over a period per population at risk; prevalence counts existing cases and equals roughly incidence times duration; relative risk comes from a cohort (incidence exposed / incidence unexposed) and the odds ratio (ad / bc) from a case-control, approximating RR only when the disease is rare.

**MNEMONIC**

### PRIDE: Prevalence Rises with Increased Duration of Existing cases

**P**revalence **R**ises with **I**ncreased **D**uration of **E**xisting cases. Because prevalence is roughly incidence times duration, longer illness inflates prevalence with incidence flat; incidence, by contrast, only ever counts the genuinely new.

**PEARL**

Match the measure to the design before you compute. A cohort hands you incidence, so you may quote a relative risk and an attributable risk. A case-control starts from cases and controls, so incidence is unavailable and you are restricted to the odds ratio, which you trust as a stand-in for relative risk only when the outcome is rare. Never quote a relative risk from a case-control study.

**GOOD TO REMEMBER**

- **Prevalence:** existing cases in a population over a time window; use it for burden and service planning; prevalence is roughly incidence times duration (P is proportional to I x d). Point prevalence = at one instant; period prevalence = over an interval.
- **Incidence:** new cases over a period per population at risk; use it for first-onset questions and to separate real rises from better survival; cumulative incidence = new cases / at-risk at start, incidence rate = new cases / person-time.
- **Relative risk:** how many times more likely disease is in the exposed; use it from a cohort; RR = incidence in exposed / incidence in unexposed; RR > 1 harmful, < 1 protective, = 1 no association.
- **Odds ratio:** the case-control measure of association; use it when incidence cannot be measured; OR = ad / bc; approximates the relative risk only when the outcome is rare.
- **Attributable risk:** the excess risk in the exposed caused by the exposure; use it to size the burden a risk factor adds; attributable risk = incidence in exposed minus incidence in unexposed; the population attributable risk scales this to the whole population and shrinks when the exposure is rare.

These measures are the shared currency of every observational and experimental design; the prevalence and incidence of individual psychiatric disorders, and the risk factors quantified by these ratios for each condition, are set out in the clinical chapters of Paper II, while the reliability and validity of the instruments used to define a case (as distinct from the frequency measures here) are covered in the Psychometry, Assessment and Descriptive Psychopathology notes.

## 17 · The National Mental Health Survey and the Indian picture

Every design and test taught earlier exists so that a number can be trusted. This topic is where those methods are spent on a single high-yield application: the National Mental Health Survey (NMHS) of India, 2015 to 2016, led by NIMHANS (Gururaj et al 2016). It is the one dataset an

examiner expects an Indian resident to quote by heart, and it is the standing answer to any viva question that begins "what do we know about the burden of mental illness in India". Read it not as a table of prevalences to memorise blindly but as the country-level payoff of the **epidemiology of** mental disorders: a cross-sectional multistage survey applying the sampling logic and cross-sectional design covered earlier in this chapter.

The examiner's real interest, and the policy headline, is not how common illness is but how few of the ill are treated. The treatment gap, the proportion of people with a diagnosable disorder who receive no care, is the number that reframes Indian psychiatry from a clinical problem into a public-health one, and it is where the marks sit. Hold the prevalences, but lead with the gap.

### 17.1 What the NMHS is, and why it is the India device

**The NMHS is India's first large multi-state national mental-health epidemiology survey.** The **national mental health survey** was conducted across 12 states of India between 2015 and 2016, coordinated by the National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, and published as NIMHANS Publication No. 129 (Gururaj et al 2016). Its purpose was to estimate the prevalence of mental disorders, describe their patterns and outcomes, and quantify the treatment gap and disability, so that mental-health policy and the District Mental Health Programme could be planned on **indian data** rather than borrowed Western figures. When a question asks for the burden of mental illness in India, this is the citation.

### 17.2 Methodology: a cross-sectional multistage survey

**The NMHS is a prevalence (cross-sectional) survey built on multistage sampling.** Methodologically the NMHS is an application of two designs taught earlier: it is a **cross-sectional** design (it measures disorder and non-disorder at one point in time, so it yields prevalence, not incidence, and cannot by itself establish that any exposure came first) delivered through a multistage random sampling frame (state, then district, then taluk or town, then village or ward, then household, then individual) so that a manageable sample can stand in for a vast and stratified population. For the anatomy of the sampling frame see the Sampling methods topic; for why a one-time snapshot gives prevalence and not causal direction see the Observational study designs topic in this chapter. Diagnoses were made with a structured diagnostic instrument (the MINI 6.0.0), the same family of fully structured interviews that anchors modern population psychiatry.

RULE **A cross-sectional survey gives prevalence, never incidence.** The NMHS tells you how many people are currently ill, not how many newly fall ill per year.

### 17.3 The headline prevalences: current and lifetime

**Roughly one Indian in ten currently has a mental disorder; nearly one in seven has had one in their lifetime.** The NMHS put the **current** prevalence of mental disorders (excluding tobacco use in most summaries) at **10.6%** (10.56% precisely) and the lifetime prevalence at **13.7%** (13.67%; Gautham et al 2020), drawn from 34,802 adults interviewed across 12 states and translating to an estimated 150 million Indians affected. Read plainly: about one adult in ten is

unwell now, and close to one in seven has been unwell at some point. These two figures are the spine of the topic; quote them together and always with their source and survey years.

| 10.6% | 13.7% |
|---|---|
| CURRENT PREVALENCE, ANY MENTAL DISORDER | LIFETIME PREVALENCE, ANY MENTAL DISORDER |

**TRAP** **Never quote a prevalence bare.** A figure without its year and its source (NMHS 2015 to 2016) is unciteable and marks nothing; pair every number with survey and year.

### 17.4 The treatment gap: the policy headline

**The Indian story is not that illness is common but that care is scarce.** The **treatment gap** is the percentage of people who meet criteria for a disorder yet receive no treatment for it. Across mental disorders in India the NMHS found the overall treatment gap to be **84.5%** (Gautham et al 2020), varying widely by disorder. This is the single most examinable NMHS fact, because it is what drives policy: it is the argument for the National Mental Health Programme, task-shifting to primary care, and closing the psychiatrist shortfall (India has fewer than 10,000 psychiatrists against a desirable minimum of about 3 per 100,000 population; Kaplan and Sadock 2024).

**RULE** **Lead with the gap, not the prevalence.** The treatment gap, not prevalence alone, is the Indian policy headline; it is what an examiner is testing.

### 17.5 How the gap varies by disorder: the substance-use picture

**The gap is widest for the substance-use and tobacco disorders.** The treatment gap is not uniform: it is largest where stigma and normalisation are greatest. For substance use disorders overall the NMHS found a treatment gap of about **91.1%**, comprising **86.3%** for alcohol use disorder, **91.8%** for tobacco use disorder, and **72.9%** for other substance use disorders (Clinical Methods in Psychiatry 2024). The lesson to carry into the clinic and the exam is that the people least likely to be in treatment are precisely those with the addictions, which reframes screening and opportunistic case-finding as the front line.

| 91.1% | 86.3% | 91.8% | 72.9% |
|---|---|---|---|
| TREATMENT GAP, SUBSTANCE USE DISORDERS | TREATMENT GAP, ALCOHOL USE DISORDER | TREATMENT GAP, TOBACCO USE DISORDER | TREATMENT GAP, OTHER SUBSTANCES |

NMHS treatment gaps for the substance-use disorders, the group in which the gap is widest (Clinical Methods in Psychiatry 2024).

### 17.6 Setting India beside the World Mental Health survey frame

**The NMHS is India's contribution to a global epidemiological project, not an isolated survey.** The World Mental Health Survey Initiative is a suite of surveys conducted throughout the world using comparable sampling and diagnostic instruments (Tasman Psychiatry 2024), carried out in representative community samples in 17 countries across Africa, the Americas, Asia, Europe and the Middle East (Kaplan and Sadock 2024). Its two enduring lessons frame the Indian numbers: mental disorders are common everywhere, and the treatment gap is substantial worldwide and worst in low-income countries. The NMHS is best read as the Indian instance of exactly this

pattern, using the same third-generation survey methodology (the ECA, the NCS and NCS-R, and the World Mental Health Survey Initiative, all built on structured instruments such as the CIDI; Tasman Psychiatry 2024). This is why "the treatment gap is global" is a defensible one-line summary and why India's overall gap of 84.5% is not an outlier but an intensification of the global picture.

**WORKED EXAMPLE: READING A TREATMENT GAP**

A treatment gap is a simple proportion. If a survey identifies 1,000 people who meet criteria for alcohol use disorder and finds that 137 of them have received any treatment for it, the treated proportion is 137 divided by 1000, which is 13.7%, so the treatment gap is 100% minus 13.7%, equals **86.3%**. That is exactly the NMHS alcohol-use-disorder figure. The arithmetic is trivial; the point of the calculation is that the gap is nothing more than one minus the proportion treated, and that a change in either the numerator (who counts as treated) or the denominator (who counts as a case) will move it. Always ask how "treated" and "case" were defined before comparing two gaps.

**COMMON TRAP**

Do not confuse the NMHS with the separate national drug-use survey. The frequently quoted figure that 14.6% of Indians aged 10 to 75 years use alcohol comes from the national survey on the magnitude of substance use in India, not from the NMHS; it is a use figure, not a disorder prevalence, and not a treatment gap. Mixing the two is a classic viva error. Keep the NMHS for disorder prevalence and treatment gap, and cite the substance-use magnitude survey separately if you quote its use figures.

**INDIA**

The whole topic is the India device for this chapter. The NMHS 2015 to 2016 (NIMHANS, 12 states) gives the numbers a resident must own: current prevalence **10.6%**, lifetime prevalence **13.7%**, and an overall treatment gap of **84.5%**, widest for the substance-use and tobacco disorders. It is the epidemiological justification for the District Mental Health Programme and for task-shifting mental-health care to primary care, given India's psychiatrist shortfall. When any Paper II clinical chapter needs an Indian prevalence for a specific disorder, the NMHS is the source to cite, but the disorder-specific prevalence must be looked up and quoted with its year, not guessed.

**PEARL**

Prevalence answers "how common", the treatment gap answers "how neglected". In India the second question carries the marks and the policy weight. A resident who can state the gap and explain why it drives task-shifting has understood the survey; one who only recites prevalences has memorised it.

MNEMONIC

**GAP: Gururaj, All-India 12 states, Prevalence 10.6 now.**

Gururaj (the lead author, Gururaj et al 2016) fixes the citation; All-India across 12 states fixes the scope; Prevalence 10.6% current (with 13.7% lifetime) fixes the headline number, so the letters of GAP also remind you that the gap itself is the point.

HOLD THIS

NMHS 2015 to 2016: current prevalence of any mental disorder 10.6%, lifetime 13.7%, and an overall treatment gap of 84.5%, which is the Indian policy headline.

GOOD TO REMEMBER

- **National Mental Health Survey (NMHS):** India's cross-sectional multistage prevalence survey (NIMHANS, 12 states, 2015 to 2016); use it for any Indian burden or treatment-gap figure; the two numbers to hold are current prevalence 10.6% and lifetime 13.7%.
- **Treatment gap:** percentage of people with a disorder who receive no care; 84.5% overall in India (widest for substance-use and tobacco); compute it as 100% minus the proportion treated.
- **World Mental Health Survey Initiative:** global suite of surveys using comparable sampling and structured diagnostic instruments (CIDI) across 17 countries; use it as the comparative frame that shows the treatment gap is global and worst in low-income settings.
- **Cross-sectional multistage design:** the NMHS method; use when you need population prevalence at one time point; remember it yields prevalence not incidence, and the one formula to know is prevalence equals cases divided by population sampled.

## 18 · Principles of screening

**Screening is prevention masquerading as a test.** **Screening** means testing apparently well people to detect disease early, before symptoms declare themselves, in the belief that earlier detection buys better outcomes (Wilson & Jungner 1968). The examiner rarely wants the mechanics of any one test; the question is whether a screening programme is justified at all, and whether an impressive-looking survival figure is real or an artefact. Two bodies of knowledge answer that: the Wilson and Jungner criteria, which decide when to screen, and the three screening biases, which explain why screening flatters itself. Master both and every screening vignette resolves cleanly (Wilson & Jungner 1968; Kaplan & Sadock 2024).

One framing distinction anchors the topic before the detail. Screening is applied to people with no complaint, so the base rate of true disease is low and the ethical bar is high: you are intervening on the well, and the harm of a false positive (anxiety, further invasive tests, overtreatment) falls on someone who was never going to be ill. That is the opposite of diagnostic testing, which is applied to the symptomatic, where prevalence is higher and the person has already sought help. This low pre-test probability is exactly why a screening test's positive predictive value collapses in the general population, a point developed in the Diagnostic test statistics topic and restated at 18.5.

### 18.1 What screening is, and how it differs from diagnosis

**Testing apparently well people to detect disease early.** Screening is the systematic application of a test to an asymptomatic population to identify those likely to have (or to be at high risk of) a disease, who are then referred for definitive diagnosis and treatment (Wilson & Jungner 1968). Two features define it. First, the target is the well, not the ill, so a screen is never itself a diagnosis: a positive screen selects people for a confirmatory diagnostic test, it does not label them with the disease. Second, the justification is downstream benefit, so a programme is only worth running if earlier detection actually changes the outcome, not merely the timing of the label. The distinction from case-finding (opportunistically testing patients already attending for another reason) and from surveillance is worth holding, but the core contrast the examiner wants is screen (asymptomatic, provisional) versus diagnose (symptomatic, definitive).

### 18.2 The Wilson and Jungner criteria (1968): the condition

**Screen only a disease worth catching early.** The classic Wilson and Jungner criteria, published for the World Health Organization in 1968, group the requirements of a sound screening programme, and the first group concerns the condition itself (Wilson & Jungner 1968). The condition must be an **important health problem** (common enough, or serious enough, to justify a population effort); it must have a **recognisable latent or early symptomatic stage** at which the screen can catch it before it declares clinically; and its **natural history must be adequately understood**, including the progression from latent to overt disease, so that we know that intercepting the latent stage genuinely alters the course. A disease with no detectable pre-clinical window, or one whose natural history is a mystery, fails at this first hurdle.

### 18.3 The Wilson and Jungner criteria: the test, the treatment and the programme

**A suitable test, an accepted treatment, and an organised, cost-effective programme.** The remaining Wilson and Jungner groups concern the test, the treatment and the programme (Wilson & Jungner 1968). The **test** must be suitable (valid, with good sensitivity and specificity), simple, safe and acceptable to the population, since a screen the public will not attend screens no one. The **treatment** must exist and be accepted: there must be an effective, agreed treatment for patients found, together with the facilities to diagnose and treat them, because detecting a disease you cannot help merely inflicts foreknowledge. The **programme** must be organised: an agreed policy on whom to treat (an agreed cut-off and referral pathway), a cost that is balanced against benefit for the whole system (cost-effectiveness), and, critically, screening understood as a continuing process, not a one-off campaign, repeated at an interval matched to the disease's natural history. Later frameworks (the UK National Screening Committee, and a 2008 WHO update) refine these, but the 1968 list remains the examinable canon.

RULE **Screen with a sensitive test, confirm with a specific one.** The first-line screen is tuned for high sensitivity to miss few true cases; every positive is then re-tested with a specific diagnostic test to weed out the false positives.

### 18.4 The sensitivity-specificity trade-off in a screening test

**A screening test is deliberately tuned for high sensitivity.** Because the cost of missing a true case (a false negative reassurance) is usually judged worse than the cost of a false alarm, the sensitivity of a screening test is set high, at the price of lower specificity and therefore more false positives (Kaplan & Sadock 2024). Those positives are not treated on the strength of the screen; they pass to a second, more **specific** diagnostic test that confirms or excludes the disease. This two-stage architecture, a sensitive sieve followed by a specific confirmation, is the standard design of a screening pathway (for example a sensitive immunoassay confirmed by a specific test, or a screening questionnaire confirmed by a structured clinical interview). Moving the cut-off trades the two errors against each other: a more lenient threshold raises sensitivity and lowers specificity; a stricter one does the reverse. The receiver operating characteristic curve, treated in its own topic, is the formal picture of that trade-off.

### 18.5 Low prevalence wrecks the positive predictive value

**In an asymptomatic population, most positives are false.** Sensitivity and specificity are intrinsic to the test and do not move with prevalence, but the positive predictive value (the chance a screen-positive person truly has the disease) is a hostage to prevalence: as prevalence falls, PPV falls (Kaplan & Sadock 2024). Screening operates precisely where prevalence is lowest, the general well population, so even an excellent test throws up a flood of false positives, and a positive screen means far less than the same result in a symptomatic clinic. The full 2x2 worked calculation (a good test whose PPV collapses from about **50%** at **10%** prevalence to roughly **8%** at **1%** prevalence) is carried in the Diagnostic test statistics topic; here, hold the consequence: a programme that looks superb in a high-risk clinic can generate mostly false alarms once rolled out to the low-prevalence general population, which is why the confirmatory specific test is not optional.

### 18.6 Lead-time bias: earlier does not mean longer

**Diagnosing earlier lengthens apparent survival without postponing death.** Lead-time bias is the spurious improvement in measured survival that arises simply because screening advances the moment of diagnosis (Kaplan & Sadock 2024). Survival is timed from diagnosis to death; if a screen detects the disease two years before it would have surfaced clinically, but the patient dies at exactly the same moment as they would have unscreened, then the survival clock has been started earlier and reads two years longer, even though no survival time was actually gained. The disease was found sooner, so the patient merely spends longer knowing they have it. Any screening series compared with a clinically-diagnosed series will show longer survival for this reason alone, before any real benefit is considered. The defence is to judge screening by mortality (deaths in the whole screened population versus the unscreened) rather than by survival-from-diagnosis, and to use a randomised trial with mortality as the endpoint.

TRAP **Improved survival in a screening series can be pure lead-time bias.** Longer survival-from-diagnosis after screening does not prove lives are saved; only a fall in population mortality does.

### 18.7 Length bias: the screen catches the indolent

**Screening preferentially catches slow, indolent disease.** Length bias (also called length-time bias) arises because a periodic screen is more likely to catch diseases that spend a long time in the detectable pre-clinical phase, and those slow-growing cases are, by their nature, the more indolent, better-prognosis ones (Kaplan & Sadock 2024). Fast, aggressive disease passes quickly from undetectable to symptomatic and often surfaces clinically in the gap between screening rounds, so it is under-represented among screen-detected cases. The screened group is therefore enriched for slow, benign disease and looks to do better, but the advantage is a sampling artefact, not a treatment effect. Length bias explains part of the apparent good prognosis of screen-detected cancers and is a cousin of lead-time bias: both make the screened group appear to fare better without any real extension of life.

### 18.8 Overdiagnosis: the extreme of length bias

**Detecting disease that would never have caused harm.** Overdiagnosis is the detection, by screening, of a genuine abnormality that met the pathological definition of disease but that would never have progressed to cause symptoms or death in the person's lifetime (Kaplan & Sadock 2024). It is the limiting case of length bias: the disease is so indolent that the patient would have died of something else first. Overdiagnosis is not misdiagnosis, the lesion is real, but the diagnosis is useless or harmful, because the person is then subjected to the anxiety, investigation and treatment of a disease that was never going to trouble them (overtreatment). It inflates apparent survival and incidence figures (more cases detected, all surviving) while doing pure harm to the individual. It is the strongest ethical argument for restraint in screening and the reason every programme must weigh benefit against the harm done to those who could never have benefited.

**WORKED EXAMPLE**

Two towns, identical disease biology. Town A is screened, Town B is not. Screen-detected patients in Town A are diagnosed on average 3 years earlier and their mean survival-from-diagnosis is 8 years; Town B's clinically-diagnosed patients survive 6 years from diagnosis. It looks like screening added 2 years. But every patient in both towns dies at the same age: the 2-year gap is exactly the lead time (3 years advanced diagnosis, offset by the different case mix), so it is lead-time bias, not benefit. On top of that, Town A's screened cases include several slow, indolent tumours that Town B's clinical presentation would have missed (they surfaced between rounds or never), so Town A's series is further enriched by length bias, and a handful of those lesions would never have harmed anyone at all (overdiagnosis). The only honest comparison is total mortality across each whole town, which shows no difference. Survival-from-diagnosis was the wrong yardstick.

### 18.9 Judging a screening programme honestly

**Mortality in a randomised trial, not survival in a case series.** Because lead-time, length and overdiagnosis all inflate survival-from-diagnosis, the correct evaluation of screening compares disease-specific (and ideally all-cause) mortality between a screened and an unscreened population, allocated at random so the two arms are otherwise alike (Kaplan & Sadock 2024). Randomisation defeats length bias (both arms get the same case mix) and the mortality endpoint

defeats lead-time bias (the clock is not reset by earlier diagnosis). The benefits of a programme are then weighed against its harms, the false positives and their cascade of tests, the overdiagnosed and overtreated, the cost, and the anxiety inflicted on the well, exactly the balance the Wilson and Jungner programme criteria demand. A screening test that improves survival figures but not population mortality has proved nothing.

**HOLD THIS**

Screen with a sensitive test and confirm with a specific one; and never trust survival-from-diagnosis in a screened series, because lead-time bias, length bias and overdiagnosis all inflate it, so judge screening only by mortality in a randomised comparison.

| BIAS | WHAT HAPPENS | WHY THE SCREENED GROUP LOOKS BETTER | DEFENCE |
|---|---|---|---|
| **Lead-time bias** | **Diagnosis advanced, death unchanged** | Survival is timed from an earlier diagnosis, so it reads longer without any life gained | Use mortality, not survival-from-diagnosis; randomised trial |
| **Length bias** | **Slow disease over-sampled** | Periodic screens preferentially catch indolent, long-pre-clinical, good-prognosis cases | Randomised allocation equalises case mix between arms |
| **Overdiagnosis** | **Non-progressive disease detected** | Real but harmless lesions are counted as cured cases, inflating incidence and survival | Weigh benefit against overtreatment harm; long-term mortality |

The three screening biases: each makes a screened series look better than it is; the hit column is the mechanism to name in a stem.

| 1968 | 3 | 8% |
|---|---|---|
| WILSON & JUNGNER YEAR | SCREENING BIASES TO NAME | PPV AT 1% PREVALENCE |

**COMMON TRAP**

Two errors the examiner rewards. First, quoting improved five-year survival as proof that a screen saves lives: that figure is exactly what lead-time and length bias inflate, and only a fall in population mortality is proof. Second, treating a positive screen as a diagnosis: a screen is a sensitive sieve, not a specific verdict, and in a low-prevalence population most of its positives are false until a specific confirmatory test says otherwise.

**MNEMONIC**

### The three biases spell L-L-O: Lead-time, Length, Overdiagnosis.

Lead-time advances the diagnosis clock; Length over-samples slow disease; Overdiagnosis catches disease that never mattered. All three inflate the screened group's apparent survival; none of them extends a single life; only population mortality tells the truth.

**INDIA**

In a low-resource, high-population setting the Wilson and Jungner programme criteria bite hardest: a screen is only defensible if the confirmatory diagnostic test and the accepted treatment are actually available and affordable at scale, and if the programme can be sustained as a continuing process rather than a one-off camp. Screening for a condition whose treatment or follow-up cannot be delivered breaches the treatment and programme criteria and merely generates anxiety and cost.

**GOOD TO REMEMBER**

- **Screening:** testing apparently well (asymptomatic) people to detect disease early; a sensitive sieve that selects people for a confirmatory diagnostic test, never a diagnosis itself; justified only if earlier detection improves the outcome, not just its timing.
- **Wilson & Jungner criteria (1968):** when to screen; four groups, the condition (important, recognisable latent stage, understood natural history), the test (suitable, acceptable), the treatment (accepted treatment plus facilities), and the programme (agreed policy on whom to treat, cost-effective, a continuing process); the WHO canon for a justified programme.
- **Screening test trade-off:** tuned for high sensitivity, accepting more false positives, then confirmed by a specific diagnostic test; the one number to hold is that PPV falls as prevalence falls, so most positives in the general population are false.
- **Lead-time bias:** earlier diagnosis lengthens apparent survival without postponing death; suspect it whenever survival-from-diagnosis improves; defeated by using mortality as the endpoint in a randomised trial.
- **Length bias:** screening preferentially catches slow, indolent, good-prognosis disease; suspect it when screen-detected cases look benign; defeated by randomised allocation equalising case mix.
- **Overdiagnosis:** detecting disease that would never have caused harm; the extreme of length bias; the strongest argument for restraint; weighed against benefit through long-term population mortality.

## Discriminate

### 19 · Discriminate: pick the design, pick the test

**This is the payoff.** Every research vignette in the examination reduces to two acts of discrimination: **which design** answers the question that was asked, and **which test** reads the data that came back. The first act is **choosing a** study design from a clinical question; the second is **test selection** from the shape of the data. The examiner rarely asks you to compute anything by hand. What is tested is whether you can move cleanly from a one-line stem to the right design, and from a described dataset to the correct statistical test, and whether you know why the wrong pick is wrong (Bland 2015; Altman 1991; Johnstone et al 2010). This topic sets out the two selectors side by side, because in practice a question flows through both in turn: the design fixes what data you will have, and that data fixes which test is legitimate.

Think of it as two interleaved lookup tables. PART A, the study-design chooser, turns a **research question** into a design, and its axis is the *nature of the question*: prevalence, a rare outcome, a rare exposure, an intervention, the totality of evidence, or meaning and experience. PART B, the

statistical-test chooser, turns the data into a test, and its axes are the *data type* (categorical versus continuous), the *number of groups* (one, two, three or more), and whether the observations are *paired or unpaired*. Hold both grids and almost every methods stem falls open.

| *data type \ comparison* | Two groups unpaired | Two groups paired | Three or more groups | Association of two variables |
|---|---|---|---|---|
| **Continuous, normal** (parametric) | **Independent t-test** | **Paired t-test** | **ANOVA** (then post-hoc) | **Pearson r** linear regression |
| **Ordinal or non-normal** (non-parametric) | **Mann-Whitney U** | **Wilcoxon** signed-rank | **Kruskal-Wallis** | **Spearman rho** |
| **Categorical counts** (proportions) | **Chi-square** (Fisher if small) | **McNemar** | **Chi-square** | **Chi-square** logistic regression |

**Fig 7.8** Choosing the right statistical test. Read down to the type of your outcome data and across to the comparison you are making, and the cell names the test. Continuous normal data take the parametric tests (t-test, ANOVA, Pearson); ordinal or non-normal data take their non-parametric twins (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Spearman); categorical counts take the chi-square family. This one grid, read with the study-design chooser, is the practical heart of the whole subject.

## 19.1 The design-choice logic: read the question, not the data

**The question dictates the design.** Before any data exist, the **design-choice** is settled by what the investigator wants to know and by what is rare or costly to observe (Bland 2015). Four sub-questions run in order. First, is this an *intervention* question (does a treatment work?) or an *observation* question (is this exposure associated with this outcome?): intervention licenses a trial, observation sends you to the observational family. Second, for observation, what is the *direction of enquiry*: from outcome back to exposure, or from exposure forward to outcome? Third, what is *rare*, the outcome or the exposure, since that decides between the two comparative designs. Fourth, does the question ask about the *current burden*, the *totality of the evidence*, or the lived *meaning* of an experience, each of which has its own home. The exact-string task here is naming **which design** fits the vignette.

RULE **Match the question to the design.** A prevalence question to cross-sectional, a rare outcome to case-control, a rare exposure or incidence question to cohort, an efficacy question to a randomised controlled trial, an evidence-synthesis question to systematic review and meta-analysis, a meaning question to qualitative.

### 19.2 PART A: the study-design chooser

**Question in, design out.** This is the first selector: read the clinical question in the left column and land on the design in the hit column (question to design). Rarity and temporality do most of the sorting; the last three rows sit apart because they answer questions of totality of evidence and of meaning rather than of individual-level association (Bland 2015; Altman 1991; Greenhalgh 2019).

| RESEARCH QUESTION (THE VIGNETTE) | DIRECTION / UNIT | DESIGN TO PICK | MEASURE IT YIELDS |
|---|---|---|---|
| **How common is the condition right now?** | Snapshot, one time point | **Cross-sectional** | Prevalence, prevalence ratio |
| **What are the risk factors for a rare outcome (or a long-latency disease)?** | Outcome to exposure (backward) | **Case-control** | Odds ratio (OR) |
| **What are the effects of a rare exposure, or the incidence of an outcome?** | Exposure to outcome (forward) | **Cohort** | Incidence; relative risk (RR) |
| **Must I prove the exposure preceded the outcome (temporality)?** | Exposure to outcome, real time | **Prospective cohort** | Incidence; relative risk (RR) |
| **Does this intervention actually work (efficacy)?** | Investigator allocates exposure | **Randomised controlled trial** | Risk ratio, absolute risk reduction, NNT |
| **What does the totality of the evidence say?** | Studies as the unit | **Systematic review and meta-analysis** | Pooled effect size (95% CI) |
| **What is the meaning or lived experience of this?** | Words, not numbers | **Qualitative** | Themes; no effect measure |
| **Only aggregate, population-level data are available?** | Group is the unit | **Ecological (mind the fallacy)** | Correlation of population rates |

Study-design chooser: read the research question, pick the design (question to design); rarity and temporality do the sorting.

TRAP **Rare exposure is not rare outcome.** A case-control study is efficient for a rare *outcome*, a cohort for a rare *exposure*; swapping them is the classic design-choice error.

### 19.3 The intervention branch: the randomised controlled trial

**When the question is efficacy.** If the vignette asks whether a treatment *works*, and the investigator can *allocate* the exposure, the answer is the randomised controlled trial (Bland 2015). Randomisation is what separates it from every observational design: allocating treatment by chance balances known and unknown confounders across arms, so a difference in outcome can be attributed to the intervention. The other pillars, allocation concealment, blinding, and intention-

to-treat analysis, protect that internal validity downstream. It is the strongest single design for a causal claim about an intervention, which is why the efficacy question routes here and nowhere else.

### 19.4 The evidence-synthesis branch: systematic review and meta-analysis

**When the question is the totality of evidence.** If the vignette asks what *all* the studies together show, the design is a systematic review, and where the studies are combined numerically, a meta-analysis (Greenhalgh 2019). A systematic review answers a focused question with a pre-specified, reproducible search and appraisal of every eligible study; the meta-analysis then pools their effect sizes into a single weighted estimate with a narrower confidence interval than any one study could give. It sits at the apex of the evidence hierarchy because it is the design whose unit of analysis is the study itself. Its validity depends wholly on the quality of the studies it pools and on whether publication bias has hidden the negative ones.

### 19.5 The meaning branch: qualitative designs

**When the question is meaning, not magnitude.** If the vignette asks about *why*, about experience, or about the meaning people attach to an illness or a service, no numerical design fits: the answer is a qualitative design (Greenhalgh 2019). Qualitative research works in words, gathered by in-depth interview, focus group or observation, and analysed by approaches such as thematic analysis, grounded theory or interpretative phenomenological analysis. It yields themes and understanding, not an effect size or a p-value, and it is judged by trustworthiness criteria (credibility, transferability, dependability, confirmability) rather than by statistical power. It is the correct pick whenever the stem asks about lived experience or meaning, and the wrong pick whenever the stem asks how many or how much.

### 19.6 The test-selection logic: read the data

**The data dictate the test.** Once a design has produced data, **test selection** turns on three facts and three facts only (Johnstone et al 2010; Altman 1991). First, the *data type*: is the outcome categorical (counts in categories) or continuous (measured on a scale), and if continuous, is it approximately normal? Second, the *number of groups* being compared: one, two, or three or more? Third, are the observations *paired* (the same subjects measured twice, or matched pairs) or *unpaired* (independent groups)? Read those three off the stem and the **statistical test** is fixed. The parametric tests (which assume an underlying normal distribution and compare means) each have a non-parametric twin (which works on ranks and assumes no distribution) for ordinal or non-normal data.

RULE **Data type x groups x pairing.** Those three coordinates alone pick the statistical test; continuous-normal goes parametric, ordinal or non-normal goes to the rank-based twin, categorical goes to chi-square.

### 19.7 PART B: the statistical-test chooser

**Data in, test out.** This is the second selector. Read across the data type, the number of groups, and the pairing to reach the correct **statistical test** in the hit column. Two independent continuous-normal groups give the independent t-test; two paired give the paired t-test; three or

more give ANOVA; their non-parametric twins are Mann-Whitney, Wilcoxon and Kruskal-Wallis; two categorical variables give chi-square; an association of two continuous variables gives Pearson (or Spearman if ordinal or non-normal) (Johnstone et al 2010; Altman 1991).

| DATA TYPE | NUMBER OF GROUPS | PAIRED OR UNPAIRED | CORRECT STATISTICAL TEST |
|---|---|---|---|
| **Continuous, normal** | One sample vs known value | n/a (single sample or a pairwise association) | **One-sample t-test** |
| **Continuous, normal** | Two | Unpaired (independent) | **Independent t-test** |
| **Continuous, normal** | Two | Paired / matched | **Paired t-test** |
| **Continuous, normal** | Three or more | Unpaired (independent) | **ANOVA (one-way)** |
| **Continuous, normal** | Three or more | Paired / repeated | **Repeated-measures ANOVA** |
| **Continuous, non-normal / ordinal** | Two | Unpaired (independent) | **Mann-Whitney U** |
| **Continuous, non-normal / ordinal** | Two | Paired / matched | **Wilcoxon signed-rank** |
| **Continuous, non-normal / ordinal** | Three or more | Unpaired (independent) | **Kruskal-Wallis** |
| **Continuous, non-normal / ordinal** | Three or more | Paired / repeated | **Friedman test** |
| **Categorical (counts)** | Two variables, two or more groups | Unpaired | **Chi-square test** |
| **Categorical (counts), small cells** | Two variables, 2x2 | Unpaired | **Fisher exact test** |
| **Categorical (counts)** | Two variables, 2x2 | Paired / before-after | **McNemar test** |
| **Two continuous variables** | Association (normal) | n/a (single sample or a pairwise association) | **Pearson correlation** |
| **Two continuous / ordinal variables** | Association (non-normal) | n/a (single sample or a pairwise association) | **Spearman rank correlation** |

Statistical-test chooser: data type by number of groups by paired/unpaired, resolving to the correct parametric test or its non-parametric twin (categorical drops to Fisher exact when any expected cell is under 5).

TRAP **Three groups is not a job for the t-test.** Running a t-test on every pair of three or more groups inflates the type I error; use ANOVA (or Kruskal-Wallis) once, then post-hoc tests.

## 19.8 The two-selector worked walk-through

WORKED EXAMPLE

**Design, then test, in one vignette.** A team asks whether a new brief psychotherapy lowers depression scores more than treatment as usual. The question is *efficacy* and the investigator can *allocate* the therapy, so PART A resolves to a randomised controlled trial. The trial randomises 120 patients to two arms and records the continuous HAM-D score at 12 weeks. Now PART B: the outcome is *continuous*, roughly *normal*, there are *two* groups, and they are *independent* (different patients in each arm), which lands on the independent t-test. Say instead each patient's HAM-D is compared before and after within the same arm: that is *two paired* measurements, so the test becomes the paired t-test. Add a third arm and compare all three endpoints at once and it becomes a one-way ANOVA. Read a 2x2 of responders vs non-responders across the two arms instead, and the counts route to a chi-square test. One vignette, both selectors, four different tests depending only on data type, group number and pairing.

## 19.9 The classic mix-ups the examiner sets

**Four errors sit at the seam of the two selectors.** Each is a discrimination the stem deliberately blurs (Bland 2015; Altman 1991).

### Odds ratio vs relative risk

A case-control study yields only an odds ratio; the relative risk needs a cohort with a denominator followed over time. The OR approximates the RR only when the outcome is rare.

### Sensitivity vs positive predictive value

Sensitivity is a property of the test (proportion of the diseased it catches) and is prevalence-independent; PPV is a property of the setting (proportion of positives who are truly diseased) and falls as prevalence falls.

### t-test where ANOVA is needed

Multiple pairwise t-tests across three or more groups inflate the false-positive rate; one ANOVA with post-hoc correction is the legitimate route.

### Correlation claimed as causation

A significant Pearson or Spearman correlation quantifies association, not cause; only a design that controls confounding and establishes temporality (ideally a trial) supports a causal claim.

COMMON TRAP

These four mix-ups recur every year. Reporting a *relative risk* from a case-control study (it gives an odds ratio); reading a high *sensitivity* as a high *positive predictive value* (they diverge as prevalence changes); running a *t-test* across three or more groups where *ANOVA* is required (multiplicity inflates type I error); and reading a *correlation* as *causation* (association is not cause). Each is a failure to discriminate at exactly the point the two selectors meet.

## 19.10 The rare-disease bridge between the two selectors

PEARL

The two selectors are linked by the rare-disease assumption. When the outcome is rare, the odds ratio from a case-control design (PART A) closely approximates the relative risk you would have obtained from a cohort, and the chi-square that tests that 2x2 table (PART B) behaves well because expected cell counts stay large enough. As the outcome becomes common, both approximations fray: the OR exaggerates the RR, and small or lopsided cells push you toward the Fisher exact test. Rarity, the same axis that steers the design choice, quietly governs which test is trustworthy on the data that design returns.

MNEMONIC

### QUESTION picks the DESIGN; DATA picks the TEST.

Two selectors, two inputs. Read the **question** (prevalence, rare outcome, rare exposure, efficacy, totality of evidence, meaning) to pick the design; read the **data** (type x number of groups x paired?) to pick the test. Never read the data to pick the design or the question to pick the test.

HOLD THIS

The question picks the design (prevalence to cross-sectional, rare outcome to case-control, rare exposure or incidence to cohort, efficacy to RCT, totality of evidence to systematic review and meta-analysis, meaning to qualitative); the data pick the test (data type x number of groups x paired/unpaired: two continuous-normal independent to independent t-test, paired to paired t-test, three or more to ANOVA, their twins Mann-Whitney / Wilcoxon / Kruskal-Wallis, two categorical to chi-square, two continuous associated to Pearson or Spearman).

**GOOD TO REMEMBER**

- **Cross-sectional:** the design for a prevalence question; use it to count current cases in a snapshot; yields prevalence = existing cases / population at that time.
- **Case-control:** the design for a rare outcome's risk factors; use it to look back from cases and controls to exposure; yields the odds ratio, OR = (odds of exposure in cases) / (odds of exposure in controls).
- **Cohort:** the design for a rare exposure's effects or for incidence; use it to follow exposed vs unexposed forward; yields incidence and relative risk, RR = (incidence in exposed) / (incidence in unexposed).
- **Randomised controlled trial:** the design for an intervention's efficacy; use it when the investigator can allocate the treatment; randomisation balances confounders and yields a risk ratio, absolute risk reduction and NNT = 1 / absolute risk reduction.
- **Systematic review and meta-analysis:** the design for the totality of evidence; use it to pool a focused question across all eligible studies; yields a single weighted pooled effect size with a 95% confidence interval.
- **Qualitative:** the design for meaning and experience; use it when the stem asks why or about lived experience; yields themes, judged by trustworthiness, with no effect measure or p-value.
- **Independent t-test:** the test for two independent continuous-normal groups; use it to compare two group means; two-sample t has n1 + n2 − 2 degrees of freedom.
- **Paired t-test:** the test for two paired continuous-normal measurements; use it for before-after or matched pairs; it tests the mean of the within-pair differences.
- **ANOVA:** the test for three or more continuous-normal group means; use it to avoid multiple t-tests; its statistic is F = MS(between) / MS(within).
- **Mann-Whitney / Wilcoxon / Kruskal-Wallis:** the non-parametric twins of the independent t-test, paired t-test and ANOVA respectively; use them on ordinal or non-normal data; they work on ranks, not means.
- **Chi-square:** the test for the association of two categorical variables; use it on absolute counts, never percentages; statistic = sum of (O − E)squared / E, dropping to the Fisher exact test when any expected cell is under **5**.
- **Pearson / Spearman:** the test for the association of two continuous variables; use Pearson when both are normal, Spearman (rank-based) when ordinal or non-normal; each returns a correlation coefficient r between −1 and +1, which is association and not causation.

For design selection and worked derivations see Bland (2015), Altman (1991) and Johnstone et al (2010); for the qualitative and evidence-synthesis branches see Greenhalgh (2019). Twin, family and adoption designs used to partition genetic and environmental variance are treated separately in the Psychiatric Genetics notes; the reliability and validity of a psychometric instrument, as distinct from the validity of a study design, are covered in the Psychometry, Assessment and Descriptive Psychopathology notes; the disorder-specific applications of these designs and tests appear in the clinical chapters of Paper II.

## Apply and consolidate

### 20 · Apply: four worked appraisal vignettes

Everything in the research-methods and biostatistics material converges on a single skill the examiner can actually test in an OSCE station or a short-answer stem: put a piece of evidence in

front of the candidate and ask them to read it correctly. This topic drills that skill through four short worked appraisal vignettes, each a compact scenario followed by the reasoning and then the read. The four moves recur across every viva: read a **forest plot**, choose the right statistical test for a described dataset, compute an index from a diagnostic **2x2 table**, and interpret a **p-value** and a **confidence interval** without over-claiming (Bland 2015, Altman 1991).

**How to use these.** Treat each vignette as a self-contained worked example: cover the read, work the reasoning yourself, then check. The arithmetic in vignette three is fully worked and verified so you can reproduce it under pressure; the other three are conceptual reads where the trap is almost always the same, over-trusting a result the data cannot support (Sackett et al 2000). Every number quoted below has been checked by hand.

## 20.1 Vignette one: read a forest plot

WORKED EXAMPLE

**Case:** A meta-analysis pools eight randomised trials of an add-on agent for depression. Each trial is a horizontal line (its point estimate and 95% confidence interval) stacked down the plot; a vertical line marks the no-effect value, and a diamond at the foot marks the pooled estimate. The pooled diamond sits to the left of the no-effect line but its right tip just crosses it. The reported heterogeneity is I-squared = **72%**.

**Reasoning.** Each study line's width shows that study's precision; a wider line is a smaller or noisier trial. The forest plot diamond is the fixed or random-effects pooled estimate, and its width is the 95% confidence interval of that pooled estimate. Whether the diamond crosses the vertical no-effect line is the whole verdict: for a ratio measure (odds ratio, relative risk) the no-effect value is **1**; for a difference measure (mean difference) it is **0**. Here the diamond crosses the null line, so the pooled effect is not statistically significant at the 0.05 level even though the central estimate favours the drug.

**The read.** A diamond that crosses the null line means the meta-analysis has not demonstrated a significant pooled effect. Then heterogeneity changes your confidence in that pooled number: I-squared of 72% is high (a common ladder reads under 25% as low, around 50% as moderate, over 75% as high), meaning most of the variation across trials is real between-study difference, not chance. High heterogeneity means the single diamond hides genuinely different underlying effects, so the pooled estimate is less trustworthy and a random-effects model with wider intervals is the honest summary (Higgins et al 2003, Bland 2015).

RULE **Cross the null, lose significance.** A forest-plot diamond touching the no-effect line (1 for ratios, 0 for differences) means a non-significant pooled result.

TRAP **Pooling despite high heterogeneity.** A tidy diamond over an I-squared of 72% is a false comfort: the trials disagree, so the single number oversimplifies.

## 20.2 Vignette two: choose the test

**WORKED EXAMPLE**

**Case:** A trial randomises patients to three treatment arms (drug A, drug B, placebo) and measures a continuous symptom score, the change on a depression rating scale, at eight weeks. The question is whether mean symptom change differs across the three arms. In this vignette the score is roughly normally distributed within each arm and variances are comparable.

**Reasoning.** Walk the chooser: the outcome is continuous, the groups are independent (different patients per arm), and there are three groups, not two. Three independent groups on a normal continuous outcome is the textbook indication for one-way ANOVA (analysis of variance), whose F statistic is the ratio of between-group to within-group variance (Johnstone et al 2010). A significant ANOVA then licenses post-hoc pairwise comparisons (Tukey or Bonferroni) to locate which arms differ.

**Why a t-test would be wrong.** A t-test compares only two means. To cover three arms you would run three separate t-tests (A vs B, A vs placebo, B vs placebo), and each carries its own **0.05** chance of a false positive, so the family-wise error inflates well above 5%. ANOVA tests all three at once at a single controlled alpha, which is exactly why it exists.

**The non-parametric twin.** If the same score were skewed or ordinal (assumptions of normality failing), swap ANOVA for its rank-based twin, the Kruskal-Wallis test, which compares three or more independent groups on medians rather than means. The two-group non-parametric equivalent (had there been only two arms) would have been Mann-Whitney U in place of the t-test.

TRAP **Serial t-tests for three arms.** Three pairwise t-tests inflate the false-positive rate; use ANOVA (or Kruskal-Wallis if non-normal) for three or more groups.

## 20.3 Vignette three: compute from a 2x2

WORKED EXAMPLE

**Case:** A brief screening questionnaire for a disorder is validated against a structured interview (the reference standard) in a clinic of 1000 attenders where the disorder prevalence is 10%. The 2x2 table comes out as below. Work out the sensitivity and the positive predictive value, then say what happens to the PPV if the same test is used where the disorder is rarer.

| SCREEN RESULT | DISORDER PRESENT | DISORDER ABSENT | ROW TOTAL |
|---|---|---|---|
| **Screen positive** | 80 (a) | 90 (b) | **170** |
| **Screen negative** | 20 (c) | 810 (d) | **830** |
| **Column total** | 100 | 900 | **1000** |

Prevalence 10%: 100 diseased, 900 non-diseased among 1000 attenders.

**Sensitivity.** Read down the disorder-present column: sensitivity = a / (a + c) = 80 / (80 + 20) = 80 / 100 = **0.80**, that is 80%. The screen detects 80% of true cases; it misses 20% (the false-negative rate).

**Positive predictive value.** Read across the screen-positive row: PPV = a / (a + b) = 80 / (80 + 90) = 80 / 170 = 0.4706, about **47%**. So even with 80% sensitivity and (from d / (b + d) = 810 / 900) 90% specificity, fewer than half of those who screen positive truly have the disorder in this clinic.

**If prevalence falls.** Sensitivity and specificity are intrinsic to the test and do not move. But re-run the same 80% sensitive, 90% specific test where prevalence is only 1%: among 10000 people, 100 are diseased and 9900 are not. True positives = 80, but false positives now = 9900 x 0.10 = 990, so PPV = 80 / (80 + 990) = 80 / 1070 = 0.0748, about **7.5%**. The very same test turns roughly nine of every ten positives into false alarms once the disorder is rare. That is why a screen that looks good in a high-risk clinic floods the general population with false positives (Wilson & Jungner 1968, Altman 1991).

| 80% | 47% | 7.5% |
|---|---|---|
| SENSITIVITY | PPV AT 10% PREVALENCE | PPV AT 1% PREVALENCE |

RULE **PPV chases prevalence.** Sensitivity and specificity stay fixed; PPV falls as the disorder becomes rarer because false positives swamp the positive row.

## 20.4 Vignette four: interpret a p-value and a CI

WORKED EXAMPLE

**Case:** A cohort study reports that an exposure is associated with a disorder at an odds ratio of 1.8 with a 95% confidence interval of 1.2 to 2.7 (p = 0.01). A parallel trial reports a mean difference in symptom score of 1.5 points (95% CI 0.2 to 2.8, p = 0.03) on a scale where clinicians regard 5 points as the smallest change a patient would notice.

**Reasoning: the odds ratio.** The point estimate 1.8 means the odds of the disorder are 1.8 times higher in the exposed. The 95% confidence interval 1.2 to 2.7 is the range of values compatible with the data; the interpretation is that if the study were repeated many times, 95% of such intervals would contain the true value. Because the interval does not include the no-effect value of **1**, the association is statistically significant, which agrees with the reported p of 0.01 (below the 0.05 threshold). For a ratio measure, crossing 1 and having p above 0.05 are the same statement.

**Reasoning: the mean difference.** Its no-effect value is **0**. The CI 0.2 to 2.8 excludes 0, so the 1.5-point difference is statistically significant (p = 0.03). But the whole interval lies below the 5-point threshold that patients can actually feel. So the result is statistically significant yet clinically trivial: a real but tiny effect, made detectable only by a large sample. Statistical significance answers "is this likely a chance finding?"; clinical significance answers "is the effect big enough to matter to a patient?", and the two can diverge in either direction (Bland 2015, Altman 1991).

**The read.** Correct the p-value too: p = 0.01 is the probability of data this extreme if the null were true, not the probability that the null is true and not the size of the effect. A narrow CI that excludes the null and clears the clinical threshold is the strongest result; a CI that excludes the null but sits below it, as here, is significant but not meaningful.

TRAP **Significant equals important.** A p under 0.05 with a CI excluding the null can still be clinically trivial if the effect sits below the minimal important difference.

## 20.5 The four appraisal moves as one habit

**One question runs through all four vignettes.** Does this evidence support the claim being made of it, or is someone over-reading it? In the forest plot the check is whether the diamond clears the null and whether heterogeneity undermines the pool; in the test-choice vignette it is whether the analysis matched the data structure or inflated error with serial comparisons; in the 2x2 it is whether a positive result means much given the prevalence; in the p-value and CI it is whether statistical significance was mistaken for clinical importance (Sackett et al 2000, Bland 2015). Learn to ask that one question and every appraisal station collapses to a short, safe read. In each worked example the vignette stem (marked case:) is deliberately terse, because the examiner wants the read, not a recital of the scenario.

HOLD THIS

Diamond crossing the null equals non-significant; three groups need ANOVA (Kruskal-Wallis if non-normal), never serial t-tests; PPV falls with prevalence while sensitivity and specificity hold; and a CI excluding the null is statistically significant but only clinically significant if the effect clears the minimal important difference.

MNEMONIC

**READ: Reach for the null line, Elect the right test, Anchor to prevalence, Distinguish significances.**

**R**each: does the forest-plot diamond (or the CI) cross the no-effect value? **E**lect: match the test to data type by group count by paired status. **A**nchor: read predictive values against the population's prevalence. **D**istinguish: statistical significance is not clinical significance.

GOOD TO REMEMBER

- **Forest plot:** the graphic summary of a meta-analysis; use it to read pooled effect and heterogeneity at a glance; the diamond crossing the null line (1 for ratios, 0 for differences) means non-significant, and I-squared over 75% is high heterogeneity.
- **I-squared (heterogeneity):** the percentage of total variation across trials due to real between-study difference rather than chance; use to judge how much to trust a pooled estimate; roughly under 25% low, around 50% moderate, over 75% high.
- **ANOVA:** the parametric test for three or more independent group means on a normal continuous outcome; use instead of multiple t-tests to control error; statistic is F = between-group variance / within-group variance.
- **Kruskal-Wallis:** the non-parametric twin of one-way ANOVA; use for three or more independent groups when normality fails; it compares ranks, so report medians.
- **Sensitivity:** the proportion of true cases the test correctly flags; use to rule out disease when negative (SnNOUT); formula = a / (a + c), read down the disease-present column and independent of prevalence.
- **Positive predictive value (PPV):** the proportion of test-positives who truly have the disorder; use to judge how much a positive result means in this population; formula = a / (a + b), and it falls as prevalence falls.
- **Odds ratio with 95% CI:** a ratio measure of association read together with its interval; use to test significance and precision at once; significant when the CI excludes the no-effect value of 1.
- **Confidence interval:** the range of values compatible with the data at 95%; use to test significance (does it cross the null?) and to gauge precision; a narrow interval excluding the null that also clears the clinical threshold is the strongest result.
- **p-value:** the probability of data at least this extreme if the null hypothesis were true; use against a pre-set threshold (conventionally 0.05); it is not the probability the null is true and not a measure of effect size.

## 21 · Consolidate: mnemonic, retrieval and synthesis

Everything in these notes reduces to one reflex the examiner is testing: an abstract lands, and you must place its **design**, judge how its **data** were analysed, and read its **disease frequency** or effect. That is the whole "THREE D's" seed planted at the start, Design then Data then Disease-frequency, and this closing topic ties the three bands into a single memorisable spine so that naming a study, choosing its test, and quoting its measure become one continuous move rather than three separate recalls (Kaplan & Sadock 2024).

**How to use this page.** Read the master mnemonic once to fix the skeleton, then close the book and work the recall prompts before you look anything up, then rebuild the whole picture as a branching map from memory. A note read passively fades in days; a note recalled and redrawn

under mild difficulty is the version that survives to the viva (Bland 2015). Consolidation is not extra content, it is the retrieval scaffold that makes the preceding content stick.

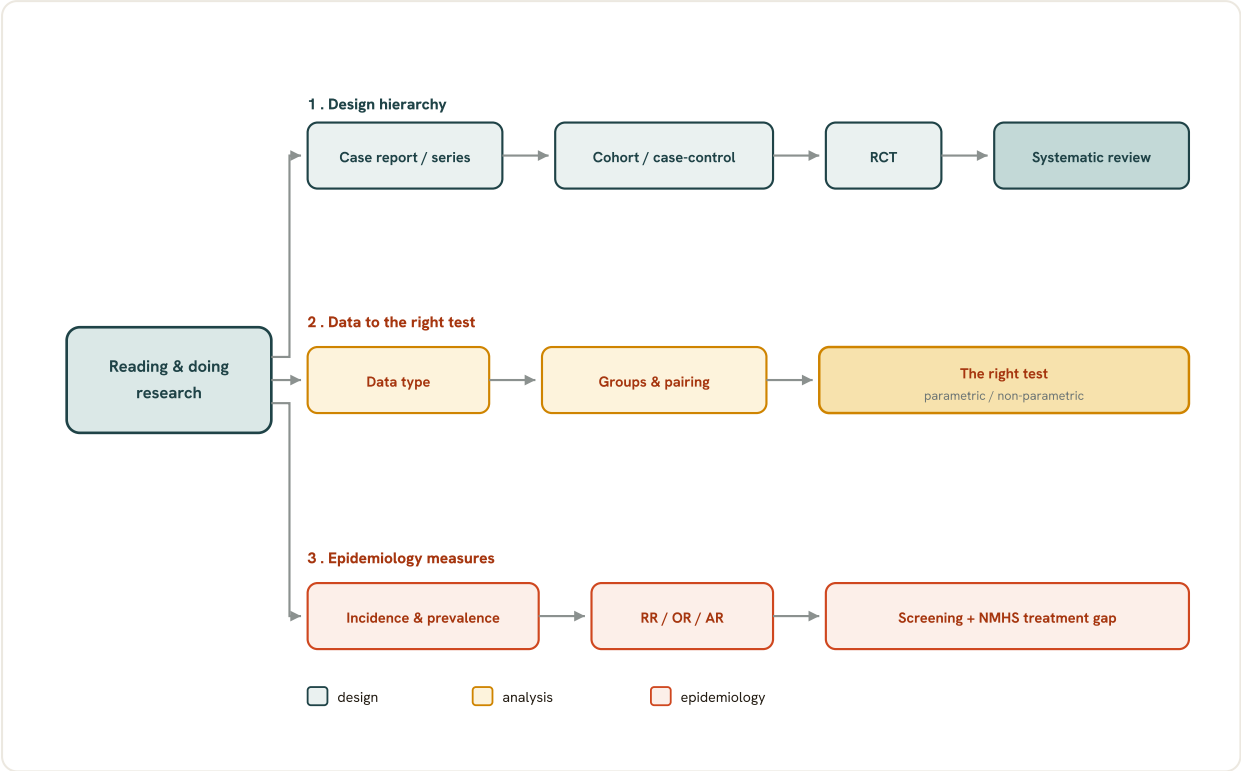



**Fig 7.9** The whole-chapter synthesis map. Three branches run out from one hub: the design hierarchy that ranks the strength of a study, the analysis path that takes a dataset from its measurement scale through the number of groups to the correct test, and the epidemiology measures that quantify how common a disorder is and how strongly an exposure raises its risk. Cover the boxes and redraw the three branches from memory as the final retrieval check.

## 21.1 The master mnemonic: one chain from design to test to measure

**The whole document on a single hook.** The three bands each get a memory anchor, and the three anchors nest under the "THREE D's" seed so that one cue unpacks the entire note. The design hierarchy runs from the weakest, most abundant evidence at the base up to the strongest, scarcest evidence at the apex; the analysis path runs from the type of data you hold to the one correct test; the epidemiology measures split into the two frequency counts and the two ratio measures of association. Learn the box below and you can reconstruct every earlier topic from its first letter.

MNEMONIC

**THREE D's -> "REACH the top, DANCE to the test, FOUR frequencies." (Design climb: Reports, Enquiry cross-sectional, Analyse case-control, Cohort, RCT, Higher-order review. Data path: Discrete-or-continuous, count of groups, Are-they-paired, Normal-or-not, then Choose. Epidemiology: Incidence, Prevalence, Relative risk, Odds ratio.)**

Unpack it band by band. **REACH** is the design climb up the evidence pyramid: **R**eports (case report and case series, the base), **E**nquiry-at-a-point (cross-sectional survey, gives prevalence), **A**nalyse-backwards (case-control, gives an odds ratio), **C**ohort (follows forward, gives incidence and relative risk), then the RCT, then the **H**igher-order synthesis (systematic review and meta-analysis, the apex): case report to cohort to RCT to systematic review, base to apex. **DANCE** is the data-to-test path: is the outcome **D**iscrete (categorical) or continuous; how many groups; **A**re they paired or unpaired; is the distribution **N**ormal (parametric) or not; then **C**hoose the cell of the test grid and **E**xecute. **FOUR frequencies** is the epidemiology quartet in reading order: **I**ncidence (new cases, from a cohort), **P**revalence (existing cases, from a cross-sectional survey), **R**elative risk (ratio of two incidences, from a cohort), **O**dds ratio (cross-product, from a case-control). One seed, three chains, the whole note: that is the payoff of the "THREE D's" you were handed at the start.

RULE **Name, test, measure, in that order.** For any abstract, say the design first, then the statistical test its data demanded, then the frequency or effect it reports; the three D's force the sequence so nothing is skipped.

## 21.2 The three chains laid side by side

**What each letter is standing in for.** A mnemonic is only worth its unpacking, so the table sets the three chains against each other: the design climb, the analysis path, and the epidemiology quartet, each with the one fact that letter must trigger. Read down any column to rehearse a band; read across to see how a single vignette moves design to test to measure.

| BAND (THE D IT SERVES) | THE CHAIN, IN ORDER | THE ONE FACT EACH LINK MUST FIRE |
|---|---|---|
| **Design climb (Design)** | Case report -> cross-sectional -> case-control -> cohort -> RCT -> systematic review | **Trustworthiness rises with control of bias; base is weak and abundant, apex is strong and scarce** |
| **Analysis path (Data)** | Data type -> number of groups -> paired or unpaired -> normal or not -> the test | **Categorical vs numerical is the top fork; parametric only when numerical and roughly normal** |
| **Epidemiology quartet (Disease-frequency)** | Incidence -> prevalence -> relative risk -> odds ratio | **Incidence and RR need a cohort; prevalence needs a cross-sectional survey; OR is the case-control stand-in** |

The three chains under the "THREE D's" seed: one column per band, read across to trace a vignette from design to test to measure.

TRAP **Quoting the wrong measure for the design.** A case-control study cannot yield a relative risk and a cross-sectional survey cannot yield an incidence; if you misname the design you will misname the measure, which is why the chain fixes them together.

## 21.3 The analysis path in one glance: data type to the right test

**The middle chain, filled in.** The DANCE letters resolve to a grid whose rows are the group-and-pairing situation and whose columns are the data type. The parametric column (interval or ratio, roughly normal) is the discriminator: pick it only when the data earn it, otherwise drop one column left to the non-parametric equivalent (Bland 2015; Altman 1991). This is the single most examined "choose the test" table, so hold it whole.

| GROUPS & PAIRING | CATEGORICAL (NOMINAL) | ORDINAL / NON-NORMAL | INTERVAL OR RATIO (PARAMETRIC) |
|---|---|---|---|
| **One sample vs known value** | Chi-square goodness-of-fit | Sign test | **One-sample t-test** |
| **Two groups, independent (unpaired)** | Chi-square test of association | Mann-Whitney U | **Independent two-sample t-test** |
| **Two groups, paired / matched** | McNemar test | Wilcoxon signed-rank | **Paired t-test** |
| **Three or more groups, independent** | Chi-square test | Kruskal-Wallis | **ANOVA (one-way), F-test** |
| **Three or more groups, repeated / paired** | Cochran's Q | Friedman test | **Repeated-measures ANOVA** |
| **Association between two variables** | Chi-square / phi coefficient | Spearman rank correlation | **Pearson correlation** |

The DANCE grid: data type across the top, group and pairing down the side, the correct test in the cell; parametric column highlighted.

## 21.4 The epidemiology quartet, worked once so the numbers stick

**The disease-frequency chain, with its arithmetic.** The two frequency counts and the two association ratios are best fixed by one 2x2 read that produces both a relative risk and an odds ratio, because seeing the two diverge is what stops you quoting one for the other. Take a cohort of 100 exposed and 100 unexposed people followed forward.

WORKED EXAMPLE

Of 100 exposed, 20 develop the disorder (incidence in exposed = 20/100 = 0.20); of 100 unexposed, 5 develop it (incidence in unexposed = 5/100 = 0.05). The relative risk = incidence in exposed / incidence in unexposed = 0.20 / 0.05 = **4.0**. The odds ratio = ad/bc = (20 x 95) / (80 x 5) = 1900/400 = **4.75**. The attributable risk = incidence in exposed minus incidence in unexposed = 0.20 minus 0.05 = **0.15**, that is 15 excess cases per 100 exposed. Note the odds ratio (4.75) sits further from 1 than the relative risk (4.0) because the disorder is common here (20 percent); make it rare and the two converge, which is the rare-disease approximation. Prevalence would come from a different design entirely, a cross-sectional snapshot, and rises with duration since prevalence is roughly incidence times duration (Kaplan & Sadock 2024).

TRAP **The odds ratio always lies further from 1 than the relative risk.** With a common outcome the OR exaggerates the RR; only when the outcome is rare do they converge, so never read a case-control OR as a relative risk when the disorder is frequent.

### 21.5 Retrieval practice: close the book and answer these first

**Test yourself before you re-read.** The evidence for durable memory is unambiguous: active recall beats re-reading, and spaced retrieval practice beats massed review (Bland 2015). So the instruction is literal, cover the answers, attempt each prompt aloud or on paper, and only then check against the earlier topics. The prompts below span all three bands, design then data then disease-frequency, and none of them prints its answer.

RETRIEVAL PRACTICE, COVER THE ANSWERS

1. Name the six rungs of the study-design hierarchy from base to apex, and state what makes the apex more trustworthy than the base.
2. Which single design feature separates the whole observational family from a randomised controlled trial?
3. A vignette gives you a rare, long-latency outcome and asks for the most efficient design: which do you pick, and which measure of association does it yield?
4. State the direction of enquiry for a cohort study versus a case-control study, and the measure each produces.
5. What is the top fork of the data-to-test path, and why does it decide whether a parametric test is even permissible?
6. Two independent groups, a continuous roughly-normal outcome: name the test. Now make the two groups paired: name the test.
7. Three or more independent groups with a numerical outcome: name the parametric test and its non-parametric equivalent.
8. Which test compares paired categorical (before-after yes/no) data in a single group?
9. Define incidence and prevalence, and give the equation that links them through duration.
10. Which study design is needed to measure incidence directly, and why can a cross-sectional survey never give it?
11. Write the formula for the relative risk and the formula for the odds ratio from a 2x2 table.
12. Under what condition does the odds ratio approximate the relative risk, and in which direction does the odds ratio err when that condition fails?
13. Distinguish attributable risk from relative risk in one line: which is a difference and which is a ratio, and what question does each answer?
14. State the "THREE D's" seed and name the memory chain that unpacks each of the three bands.
15. For any abstract, what is the fixed order in which you name the design, the test, and the measure?

### 21.6 The synthesis map: redraw the whole note from memory

**Rebuild the three branches on a blank page.** The final consolidation move is to draw a synthesis map, a single branching diagram you redraw from memory with the "THREE D's" seed at the

centre and three branches radiating out: the *design hierarchy* climbing case report to cross-sectional to case-control to cohort to RCT to systematic review; the *data-to-test analysis path* forking on data type, group number, pairing and normality down to the named test; and the *epidemiology measures* splitting into the two frequency counts (incidence, prevalence) and the two association ratios (relative risk, odds ratio). If you can redraw the synthesis map with every branch labelled and every leaf carrying its one fact, without the notes open, the note has consolidated; a gap on the map is exactly the topic to revisit.

**GOOD TO REMEMBER**

- **Design hierarchy:** ranks study designs by control of bias; use it to place any abstract and know what its rung can answer; order (base to apex) = case report -> cross-sectional -> case-control -> cohort -> RCT -> systematic review.
- **Data-to-test path:** selects the correct statistical test; use it once the data are in hand; the deciding chain = data type, number of groups, paired vs unpaired, normal vs not, then the test cell.
- **Incidence:** new cases over a period per population at risk; use it for a first-onset question; needs a cohort, and it is the numerator of the relative risk.
- **Prevalence:** existing cases at a point or over an interval; use it for burden and service planning; from a cross-sectional survey, and prevalence is roughly incidence times duration.
- **Relative risk:** how many times more likely disease is in the exposed; use it from a cohort; RR = incidence in exposed / incidence in unexposed (worked value 4.0).
- **Odds ratio:** the case-control measure when incidence cannot be measured; use it from a case-control study; OR = ad/bc (worked value 4.75), approximating RR only when the outcome is rare.

**HOLD THIS**

The "THREE D's" seed unpacks the whole note: name the Design on the hierarchy, choose the Data's test by type-groups-pairing-normality, then quote the Disease-frequency measure that design can legitimately give, in that order, every time.

## Previously asked

Research methodology, biostatistics and epidemiology sit in the first paper's basic-sciences territory. In the theory catalogue sampled here they have not appeared as standalone long essays or short notes: the subject is examined **applied** rather than recited, folded into clinical questions (the epidemiology and risk factors of a named disorder, the interpretation of a trial in a management answer, the ethics of consent in a forensic scenario) and, above all, in the critical-appraisal and thesis-protocol components. Genuine standalone items will be added here as further question catalogues are sourced, never invented.

**? HOW THIS TERRITORY IS ACTUALLY TESTED**

Q As an **appraisal task**: given a study or a forest plot, name the design, its measure of association, its main bias, and whether the conclusion is justified.

Q As an **applied number**: read a 2x2 for sensitivity and predictive value, interpret a p-value and a confidence interval, or compute a relative risk or an NNT.

- Q As an **Indian-data prompt**: quote the National Mental Health Survey prevalence and the treatment gap to frame a public-health answer.

- Q As **research ethics**: the four principles, informed consent and the ethics committee within a scenario, most often in the forensic and community papers.

Prepare this material to be used, not just remembered. The two selector tables (choosing a design, choosing a test) and the four worked appraisal cases in this note are deliberately built to that examinable skill. The disorder-specific epidemiology that recurs in clinical essays is developed in the clinical chapters of Paper II; the professional and forensic ethics that the ethics questions usually target are developed in the mental health legislation and forensic psychiatry notes; this note owns the general research toolkit, the statistics, and the national survey numbers.

# Quick review

| The whole subject in one table | |
|---|---|
| EVIDENCE-BASED MEDICINE | The five steps (Ask, Acquire, Appraise, Apply, Assess); the focused question via PICO; the null and alternative hypotheses. The evidence pyramid ranks designs by control of bias: case report, cross-sectional, case-control, cohort, randomised controlled trial, then systematic review and meta-analysis at the apex. |
| OBSERVATIONAL DESIGNS | Cross-sectional is a snapshot giving prevalence. Case-control traces outcome back to exposure, gives the odds ratio, and is efficient for rare outcomes. Cohort follows exposure forward to outcome, gives incidence and the relative risk, and suits rare exposures. Ecological studies compare populations (mind the ecological fallacy). |
| THE RANDOMISED CONTROLLED TRIAL | Randomisation balances known and unknown confounders; allocation concealment hides the next assignment; blinding hides the treatment; intention-to-treat analyses everyone as randomised. Parallel, crossover, factorial and cluster designs. Drug-trial phases I (safety) to IV (post-marketing). |
| VALIDITY, BIAS, CONFOUNDING | Internal validity (true here) versus external validity (generalisable). Bias is systematic error: selection and information (including recall) bias. Confounding is a third variable on neither causal path, controlled at design (randomisation, restriction, matching) or analysis (stratification, regression). Distinct from psychometric reliability and validity. |
| SAMPLING | Probability sampling (simple random, systematic, stratified, cluster, multistage) permits generalisation; non-probability sampling (convenience, purposive, snowball) does not. Stratify to guarantee a small subgroup; clustering adds a design effect. |
| QUALITATIVE METHODS | Grounded theory, thematic analysis, phenomenology, ethnography; the focus group and the in-depth interview. Rigour through saturation, triangulation, reflexivity and an audit trail. For meaning and hypothesis generation, not statistical generalisation. |
| SYSTEMATIC REVIEW AND META-ANALYSIS | The reproducible protocol-driven review versus the statistical pooling that may sit inside it. Read the forest plot (point estimates, confidence intervals, the null line, the pooled diamond). Heterogeneity via I-squared; publication bias via the funnel plot; certainty via GRADE. |
| REPORTING GUIDELINES AND ETHICS | CONSORT for trials, PRISMA for reviews, STROBE for observational studies, GRADE for certainty. The four principles (autonomy, beneficence, non-maleficence, justice), Helsinki and Nuremberg, informed consent, the ethics committee, and the ICMR 2017 guidelines in India. |
| DATA TYPES AND SCALES | Nominal, ordinal, interval and ratio; categorical versus continuous (discrete or continuous). The measurement scale decides the legal summary |

| The whole subject in one table | |
|---|---|
| | statistic and the legal test, so it is the first thing to settle. |
| DESCRIBING DATA AND THE NORMAL CURVE | Central tendency (mean, median, mode) and dispersion (range, interquartile range, variance, standard deviation). The normal distribution: mean equals median equals mode, and the empirical rule that 68, 95 and 99.7 percent lie within one, two and three SDs. Skew and kurtosis; the z-score. |
| TESTS OF SIGNIFICANCE | Parametric tests for normal interval data: the t-test (one-sample, independent, paired), ANOVA with post-hoc for three or more groups. Non-parametric twins for ordinal or non-normal data: Mann-Whitney, Wilcoxon, Kruskal-Wallis; chi-square (Fisher exact for small cells) for categorical data. |
| CORRELATION, REGRESSION, FACTOR ANALYSIS | Pearson for linear association of continuous normal variables, Spearman for ranks. Linear regression predicts a continuous outcome (r-squared as variance explained); logistic regression predicts a binary one (adjusted odds ratios). Factor analysis reduces items to latent factors (exploratory versus confirmatory, eigenvalues, loadings, PCA). |
| DIAGNOSTIC TEST STATISTICS | From the 2x2: sensitivity and specificity are intrinsic (SnNOUT, SpPIN); the positive and negative predictive values are what the clinician wants but move with prevalence; the likelihood ratio shifts pre-test to post-test odds. At low prevalence a positive screening test is often a false alarm. |
| STATISTICAL INFERENCE | The p-value is the probability of the data (or more extreme) under the null; alpha is set at 0.05. The confidence interval gives the plausible range (crossing 0 for a difference, or 1 for a ratio, means non-significant). Type I error is a false positive (alpha), type II a false negative (beta); power is 1 minus beta, aim 80 percent. Sample size and effect size drive it. |
| SURVIVAL, NNT AND ROC | Kaplan-Meier plots event-free probability over time with censoring; the log-rank test compares curves and the hazard ratio quantifies the difference. NNT is one over the absolute risk reduction. The ROC curve plots sensitivity against one minus specificity, and the AUC grades discrimination (0.5 chance, 1.0 perfect). |
| DISEASE FREQUENCY AND ASSOCIATION | Incidence (new cases) versus prevalence (existing cases, point or period; roughly incidence times duration). Relative risk from a cohort, the odds ratio from a case-control (it approximates the RR when the outcome is rare), and the attributable risk as the excess in the exposed. |
| THE INDIAN PICTURE (NMHS) | The National Mental Health Survey 2015-16 gives India's current and lifetime prevalence and, its policy headline, the large treatment gap that varies by disorder. Read it beside the world mental health frame; it is an applied cross-sectional multistage survey. Every quoted number is verified against the source. |

| The whole subject in one table | |
|---|---|
| SCREENING | The Wilson and Jungner criteria (the condition, the test, the treatment, the programme). Lead-time bias (earlier diagnosis, not longer life), length bias (indolent disease over-sampled) and overdiagnosis. Screen with a sensitive test and confirm with a specific one; low prevalence lowers the predictive value. |
| DISCRIMINATE | Two selectors: the study-design chooser takes a question to a design (prevalence to cross-sectional, rare outcome to case-control, rare exposure or incidence to cohort, efficacy to RCT, totality to systematic review). The statistical-test chooser takes the data type, the number of groups and pairing to the correct parametric or non-parametric test. |
| APPLY AND CONSOLIDATE | Four worked appraisals: read a forest plot, choose the test for a dataset, compute sensitivity and PPV from a 2x2, and interpret a p-value with its confidence interval. Then the master mnemonic (design to data to disease frequency), answer-free retrieval prompts, and the three-branch synthesis map to redraw from memory. |

## References

The material is grounded in the standard texts (the source hierarchy below); statistical formulae and definitions are drawn from the medical-statistics references, and each landmark, contested or numeric claim carries a PubMed-verified journal citation. Every PMID here was checked against PubMed this build, and any citation whose identifier did not resolve to the cited paper was corrected or dropped. The National Mental Health Survey figures are taken from the primary results paper, not from a textbook summary.

### SOURCE HIERARCHY (TEXTBOOKS AND STATISTICS REFERENCES)


1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall/CRC; 1991. [the medical-statistics depth bar: distributions, tests, sample size, diagnostic statistics]
2. Bland M. An Introduction to Medical Statistics. 4th ed. Oxford: Oxford University Press; 2015. [describing data, significance testing, correlation and regression]
3. Johnstone EC, Owens DGC, Lawrie SM, McIntosh AM, Sharpe M, eds. Companion to Psychiatric Studies. 8th ed. Edinburgh: Churchill Livingstone/Elsevier; 2010. [the psychiatric research-methods and biostatistics chapter, the primary parsed source]
4. Geddes JR, Andreasen NC, Goodwin GM, eds. New Oxford Textbook of Psychiatry. 3rd ed. Oxford: Oxford University Press; 2020. [evidence-based medicine and clinical epidemiology]
5. Boland R, Verduin ML, eds. Kaplan & Sadock's Comprehensive Textbook of Psychiatry. 11th ed. Philadelphia: Wolters Kluwer; 2024. [the depth bar; psychiatric epidemiology and statistics]
6. Tasman A, Kay J, Lieberman JA, First MB, Riba MB, eds. Tasman's Psychiatry. Chichester: Wiley-Blackwell. [psychiatric epidemiology; the World Mental Health Survey Initiative]
7. Clinical Methods in Psychiatry. 1st ed. New Delhi: AIIMS Publication; 2024. [the mental-health survey figures and the Indian context]
8. Greenhalgh T. How to Read a Paper: The Basics of Evidence-Based Medicine and Healthcare. 6th ed. Oxford: Wiley-Blackwell; 2019. [critical appraisal, reading the literature]


### FOUNDATIONAL SOURCES

1. Wilson JMG, Jungner G. Principles and Practice of Screening for Disease. WHO Public Health Papers No. 34. Geneva: World Health Organization; 1968. [the ten screening criteria]
2. Cochrane AL. Effectiveness and Efficiency: Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust; 1972. [the case for the randomised controlled trial and evidence-based care]
3. Higgins JPT, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 6. London: Cochrane; 2023. [systematic-review method, heterogeneity and the I-squared interpretation bands]
4. World Medical Association. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Adopted 1964, current revision 2013. [the research-ethics reference standard]
5. Indian Council of Medical Research. National Ethical Guidelines for Biomedical and Health Research Involving Human Participants. New Delhi: ICMR; 2017. [the Indian research-ethics framework]

## JOURNAL REFERENCES (PUBMED-VERIFIED, 20)

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2. PMID: 8555924.
2. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268(17):2420-5. PMID: 1404801.
3. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. N Engl J Med. 1983;309(22):1358-61. PMID: 6633598.
4. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408-12. PMID: 7823387.
5. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60. PMID: 12958120.
6. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34. PMID: 9310563.
7. Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. Lancet. 2004;363(9418):1341-5. PMID: 15110490.
8. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. PMID: 20332509.
9. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535. PMID: 19622551.
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. Lancet. 2007;370(9596):1453-7. PMID: 18064739.
11. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6. PMID: 18436948.
12. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting (MOOSE). JAMA. 2000;283(15):2008-12. PMID: 10789670.
13. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283(20):2701-11. PMID: 10819955.
14. Appelbaum PS. Assessment of patients' competence to consent to treatment. N Engl J Med. 2007;357(18):1834-40. PMID: 17978292.
15. Stevens SS. On the theory of scales of measurement. Science. 1946;103(2684):677-80. PMID: 17750512.
16. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-8. PMID: 9553006.
17. Sterne JAC, Davey Smith G. Sifting the evidence, what's wrong with significance tests? BMJ. 2001;322(7280):226-31. PMID: 11159626.
18. Fagan TJ. Letter: Nomogram for Bayes's theorem. N Engl J Med. 1975;293(5):257. PMID: 1143310.
19. Paykel ES, Scott J, Teasdale JD, et al. Prevention of relapse in residual depression by cognitive therapy: a controlled trial. Arch Gen Psychiatry. 1999;56(9):829-35. PMID: 12884889.
20. Gautham MS, Gururaj G, Varghese M, et al. The National Mental Health Survey of India (2016): Prevalence, socio-demographic correlates and treatment gap of mental morbidity. Int J Soc Psychiatry. 2020;66(4):361-72. PMID: 32126902.

# Index


2x2 table 3
Absolute risk reduction 10
active recall 120
Alcohol use disorder 97
Allocation concealment 3
alpha 65
Alternative hypothesis 6
analysis of variance 64
ANOVA 16
Assent 50
Attributable risk 93
AUC 49
Audit trail 39
Autonomy 49
Belmont Report 49
Beneficence 49
Berkson's bias 25
beta 79
Bias 3
Blinding 3
Bonferroni 64
Capacity 50
Case report 3
Case-control 5
Case-control study 5
Case-finding 97
Categorical data 54
Censoring 85
chi-square 16
Chi-square test 44
Cluster randomised 20
Cluster sampling 30
Cochrane 43
coefficient of determination 68
Cohen's d 82
Cohort 3
Cohort study 5
confidence interval 3
Confirmability 38
confirmatory factor analysis 71
Confounder 7
Confounding 3
CONSORT 48
Constant comparison 37
Content analysis 37
Continuous data 54
Convenience 30
correlation 3
Cox regression 27
Credibility 38
Critical appraisal 47
Cross-sectional 5
Cross-sectional study 10
Cross-sectional survey 5
Crossover 20
Cumulative incidence 90
data-to-test path 118
Declaration of Helsinki 49
Dependability 38
Design effect 32
Diamond 41
Discrete data 54
Drug trial 21
Ecological fallacy 11
Ecological study 11
effect size 6
Egger's test 45
eigenvalue 71
Empirical rule 58
Ethics committee 47
Ethnography 37
Evidence pyramid 4
Evidence-based medicine 4
exploratory factor analysis 71
Exposure 5
External validity 24
F-test 64
factor analysis 67
factor loading 71
Factorial 20
Fagan's nomogram 77
Fisher exact 65
Fisher exact test 65
Fixed-effect 44
Focus group 38
Forest plot 3
Funnel plot 4
Gaussian 60
Generalisability 24
GRADE 7

Grounded theory 37
Hazard ratio 70
Healthy worker effect 25
Heterogeneity 8
Hierarchy of evidence 4
Hypothesis 6
I-squared 16
ICMR National Ethical Guidelines 2017 51
In-depth interview 38
Incidence 3
Incidence rate 14
Independent t-test 64
Information bias 24
Informed consent 47
Institutional review board 51
Intention-to-treat 3
Internal validity 24
Interquartile range 53
Interval scale 52
Jungner 77
Justice 49
Kaiser criterion 71
Kaplan-Meier 85
Kruskal-Wallis 16
Kurtosis 61
Lead-time bias 101
Length bias 102
Levels of evidence 7
Likelihood ratio 74
Likert item 53
linear regression 69
Log-rank test 85
logistic regression 70
Mann-Whitney U 16
Mann-Whitney U test 53
Mean 7
Measure of association 12
Median 52
Median survival time 86
Member checking 39
Meta-analysis 3
minimum clinically important difference 82
Misclassification 25
mnemonic 8
Mode 7
MOOSE 48
Multistage 30
Multistage sampling 33
multivariable regression 70
National Mental Health Survey 3
Negative predictive value 74
Nested case-control study 13
NIMHANS 95
Nominal scale 52
Non-maleficence 49
non-parametric 3
Non-probability sampling 30
Normal distribution 57
Null hypothesis 6
Number needed to harm 87
Number needed to treat 10
Nuremberg Code 49
Odds ratio 3
Ordinal scale 52
Overdiagnosis 102
Oxford CEBM 7
p-value 3
paired t-test 16
Parallel-group 20
parametric 3
Participant observation 37
Pearson correlation 56
Per-protocol analysis 20
Period prevalence 90
Person-time 90
Phenomenological approach 37
PICO 5
Point prevalence 9
Population attributable risk 93
Positive predictive value 74
post-hoc 64
Post-test probability 77
power 8
power calculation 82
Pre-test probability 77
Prevalence 3
principal component analysis 71
PRISMA 45
Probability 6
Probability sampling 30
Prospective cohort 14
Publication bias 7
Purposive 30
Qualitative 3
Random-effects 44
Randomisation 3

Randomised controlled trial 3
Range 45
Ratio scale 52
Recall bias 13
Reflexivity 39
Relative risk 3
Reliability 3
Research ethics 3
Research question 6
retrieval practice 120
ROC curve 75
sample size 20
Sampling bias 34
Sampling error 34
Sampling frame 30
Saturation 37
scree plot 71
Screening 3
Selection bias 13
Sensitivity 73
Sensitivity of a screening test 101
significance 3
Simple random 18
Skew 54
SnNOUT 74
Snowball 30
Spearman rank correlation 56
Specificity 73
SpPIN 74
Standard deviation 53
Standard error of the mean 59
Standard normal 60
Standardisation 27
Stratification 11
Stratified 18
STROBE 48
student t-test 63
Study-design hierarchy 4
Surrogate consent 50
Survival analysis 84
synthesis map 117
Systematic review 3
Systematic sampling 30
Temperature in Celsius 52
Thematic analysis 37
Tobacco use disorder 97
Transferability 37
Treatment gap 12
Triangulation 39
Trustworthiness (Lincoln and Guba) 41
Tukey 64
Tumour stage 52
type i error 64
type ii error 64
Variance 32
Wilcoxon signed-rank 16
Wilson 77
Wilson and Jungner criteria 99
World Mental Health Survey Initiative 97
Z-score 60

## PRINTING AND DISTRIBUTION

Notes Next Door (NND), Delhi is the official printing and distribution partner for Weave Sprint. The soft-copy PDFs and the complete print-ready set are always available at weave.clinic/sprint, should you prefer to print the notes yourself. Weave Sprint is open source and free, and as part of Weave's commitment to residents it will stay that way, forever. We do, however, strongly recommend a printed set from Notes Next Door: in our experience, they have been the most reliable printing partner we have worked with.