

Topic 12

University Medicine

What this course is, and how to use it

This is the twelfth course in University Medicine: **epidemiology**, **biostatistics** and **evidence-based medicine** (EBM) – the toolkit that turns clinical questions into numbers and numbers into decisions. It aligns with the reviewed UNI-BUS-05 statistics material where the mathematics overlaps.

Scope limit, stated once: a study course, not a medical degree; every dataset below is synthetic or openly licensed, every abstract appraised is fictional, and nothing here licenses or guides treatment.

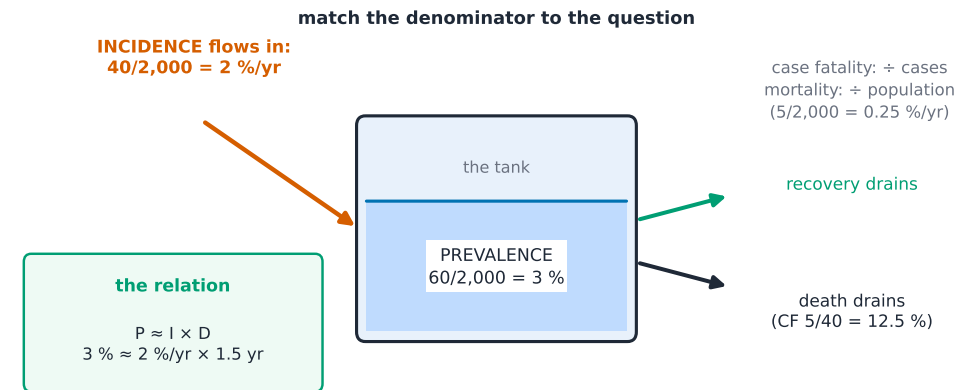
- Units 12.1–12.3 measure and compare: disease frequency, study designs, bias and causation.
- Units 12.4–12.6 quantify: diagnostic tests, inference, trials and reviews.
- Unit 12.7 applies: appraisal, certainty, and the patient in front of you.

The EBM habit: **ask in PICO, measure in risks and ratios, decide with intervals not rituals** – every calculation below is synthetic, machine-checked, and something you recompute.

Measuring disease: incidence and prevalence

Why it matters. Before any comparison, disease must be counted – and the two classic counts answer different questions: how fast new cases appear, and how many exist at once.

The idea, with a picture. **Incidence** = new cases ÷ people at risk ÷ time; **prevalence** = existing cases ÷ population at an instant ✓. They are linked: prevalence ≈ incidence × duration – long diseases pile up, brief ones vanish between surveys.



Incidence fills the tank; cure and mortality drain it – prevalence is the water level

The derivation (why standardisation exists). Crude rates mix two effects: disease risk and population age structure. A district with more older residents can show a higher crude rate with LOWER age-specific rates ✓. Direct standardisation applies each group's age-specific rates to one standard population and sums – the artificial total is comparable because the age mix is now identical.

Worked example 1 – a synthetic town. 2,000 residents, one year, 40 new cases of a disease → incidence $40/2,000 = 2 \%$ per year ✓. On survey day 60 people have the disease → point prevalence $60/2,000 = 3 \%$ ✓. Check the link: $3 \% \approx 2 \%/ \text{year} \times 1.5 \text{ years average duration}$ – the numbers are a consistent system.

Worked example 2 – clinical link: mortality's three faces. (Fictional dataset) of the 40 new cases, 5 died within the year: **case fatality** $5/40 = 12.5 \%$ ✓; the yearly deaths from this cause in the whole town were $5/2,000 = 0.25 \%$ (cause-specific mortality rate); and if the town recorded 20 deaths from all causes that year, this disease caused $5/20 = 25 \%$ of them (proportional mortality). Same deaths, three denominators, three different questions.

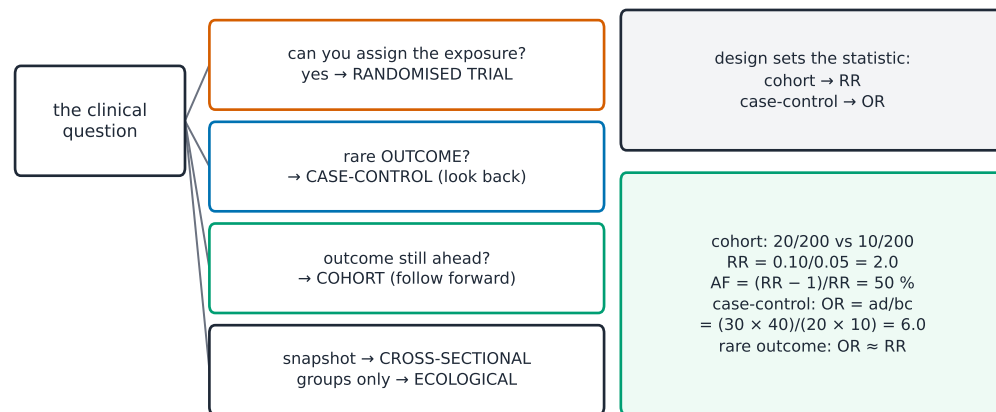
Common mistakes. Calling prevalence a rate (it is a proportion – no time unit). Comparing crude rates across populations with different age structures (standardise first). Reading case fatality as mortality (the denominator is cases, not people).

Summary. Incidence = new/at-risk/time; prevalence = existing/population; $P \approx I \times D$; standardisation fixes the age mix; case fatality (cases), cause-specific mortality (population), proportional mortality (all deaths) – match the denominator to the question.

Study designs

Why it matters. Every clinical question has designs that answer it well and designs that cannot – choosing the design is choosing which biases you will fight.

The idea, with a picture. The ladder: randomised trial (experiment) → cohort (follow exposed and unexposed forward) → case-control (look backward from outcome) → cross-sectional (one snapshot) → ecological (groups, not individuals) ✓.



Rare disease or outcome → case-control; exposure measured ahead of a common outcome → cohort; both questions at once → trial

The derivation (why RR belongs to cohorts and OR to case-control). A cohort samples by EXPOSURE, so risks in each arm are estimable: **relative risk**

$RR = \frac{a/(a+b)}{c/(c+d)}$ ✓. A case-control samples by OUTCOME, so the “risk” in each

arm is an artefact of the sampling ratio – only the **odds ratio** $OR = \frac{a/c}{b/d} = \frac{ad}{bc}$ is valid ✓. The design dictates the statistic, not preference.

Worked example 1 – one synthetic cohort. Exposed 200, 20 events; unexposed 200, 10 events: risks 0.10 vs 0.05 → $RR = 2.0$ ✓, attributable fraction among exposed $(RR - 1)/RR = 50\%$ ✓ – the exposed group’s events are half “because of” the exposure under the assumptions.

Worked example 2 – clinical link: the case-control mirror. (Fictional) 30 of 40 cases exposed; 20 of 60 controls exposed: $OR = \frac{30 \times 40}{10 \times 20} = 6.0$ ✓. A rare outcome makes OR approximate RR – one number, two designs, and the sampling decides which is legal.

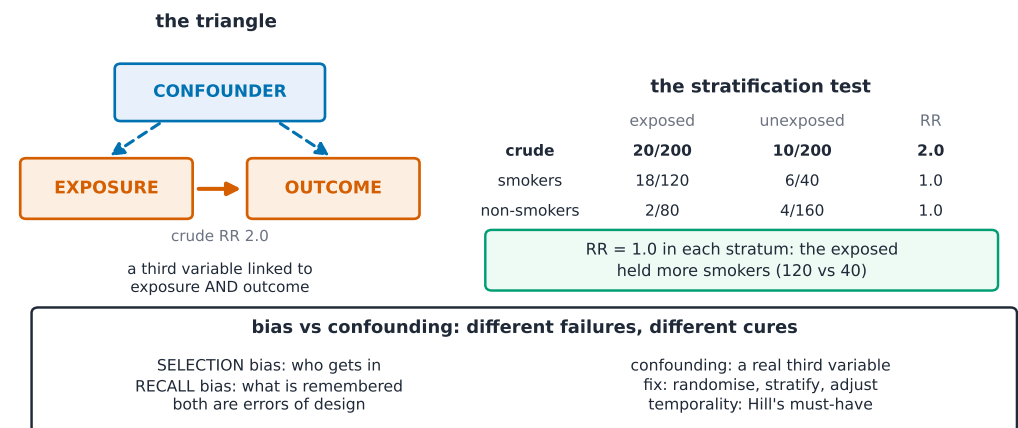
Common mistakes. Computing RR from a case-control study (the denominators are chosen by the researcher, not found). Reading a cross-sectional odds ratio as causal (exposure and outcome are simultaneous). Equating “prospective” with “better” (a prospective cohort still confounds; only randomisation balances the unknowns).

Summary. Trial → cohort (RR) → case-control (OR) → cross-sectional → ecological; $RR = \frac{a/(a+b)}{c/(c+d)}$, $OR = ad/bc$; cohort samples by exposure, case-control by outcome; rare-disease $OR \approx RR$; attributable fraction $(RR - 1)/RR$.

Bias, confounding and causation

Why it matters. Numbers lie in three ways: systematically (bias), through a third variable (confounding), and by association without cause. Naming the failure mode is half of critical appraisal.

The idea, with a picture. **Bias** is systematic error in design or measurement: **selection bias** (who gets in), **information/recall bias** (what is remembered or recorded). **Confounding** is a third variable linked to both exposure and outcome that pretends to be the cause ✓.



The confounder sits on the back path – stratify and the spurious association collapses

The derivation (why stratification is the proof). If a third variable is the true driver, the association should vanish inside strata where it is constant: crude RR 2.0, but within non-smokers RR 1.0 and within smokers RR 1.0 → smoking was the confounder ✓. If the association persists within strata (say 2.0 and 1.9), the exposure keeps its own signal. Randomisation prevents confounding prospectively by balancing even the unknowns ✓.

Worked example 1 – the stratified table. (Synthetic) crude: exposed 20/200 events vs unexposed 10/200 → RR 2.0. Stratify by smoking: non-smokers 6/100 vs 3/100 (RR 2.0 ✓ inside), smokers 14/100 vs 7/100 (RR 2.0) – the association survives **stratification** : smoking was NOT the confounder here; the exposure’s signal stands. Now give the same crude table a different split – exposed: 120 smokers (18 events) and 80 non-smokers (2 events); unexposed: 40 smokers (6 events) and 160 non-smokers (4 events). Smokers’ risk is 15 % in both arms and non-smokers’ 2.5 % in both, so RR is 1.0 inside each stratum: the exposed group simply held more smokers, and smoking WAS the confounder ✓. One crude table, two verdicts – only the strata can tell them apart.

Worked example 2 – clinical link: the appraisal questions. (Fictional abstract) “coffee drinkers had twice the risk”: ask who was studied (selection), how exposure was measured (recall – patients with illness remember differently), and what a confounder would need to look like (associated with coffee AND with the outcome, like smoking). Bradford Hill’s viewpoints – temporality first, then strength, consistency, dose-response, plausibility – turn a headline into a checklist ✓.

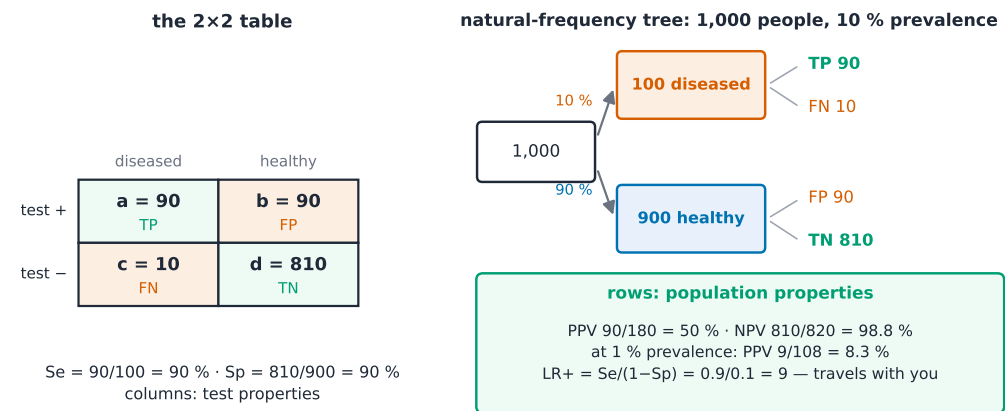
Common mistakes. Treating confounding and measurement bias as one problem (confounding is a mixing of effects through a third variable, cured by randomisation, stratification or adjustment; selection and information bias are errors in who is studied and how things are measured, which no analysis fully repairs). Expecting big data to cure confounding (more precision around a biased estimate). Demanding every Hill viewpoint (they are viewpoints, not a scorecard; temporality is the only non-negotiable).

Summary. Bias = systematic error (selection, information/recall); confounder = third variable on both paths; the stratification test collapses the association if confounded; randomisation balances the unknowns; Hill’s viewpoints with temporality first.

Diagnostic tests

Why it matters. A test’s numbers are not properties of the test alone – sensitivity and specificity belong to the test, but **predictive values** belong to the population using it. Misreading this is the commonest quantitative error in clinical reasoning.

The idea, with a picture. The 2×2: **sensitivity** = $a/(a+c)$ (diseased, test positive); **specificity** = $d/(b+d)$ (healthy, test negative); **PPV** = $a/(a+b)$, **NPV** = $d/(c+d)$ – and the natural-frequency tree shows why prevalence moves the predictive values ✓.



Sensitivity and specificity live in the columns; predictive values live in the rows – prevalence sets the row widths

The derivation (why prevalence drives PPV). $PPV = \frac{Se \times P}{Se \times P + (1-Sp)(1-P)}$ ✓. With Se = Sp = 90 %: at prevalence 10 %, $PPV = \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.1 \times 0.9} = 0.5$ ✓ – even an excellent test is a coin-flip in a low-prevalence clinic. The false-positive arm scales with the healthy majority.

Worked example 1 – the natural-frequency tree (synthetic). 1,000 people, prevalence 10 %: 100 diseased → 90 true positives, 10 false negatives; 900 healthy → 810 true negatives, 90 false positives ✓. PPV = 90/180 = 50 %; NPV = 810/820 = 98.8 % ✓. Same test at prevalence 1 %: 9 TP vs 99 FP → PPV 9/108 = 8.3 % ✓ – one tree, two worlds.

Worked example 2 – clinical link: the likelihood ratio travels. $LR+ = Se/(1-Sp) = 0.9/0.1 = 9$ ✓: a positive result multiplies the pre-test odds by 9 whatever the clinic’s prevalence. Pre-test odds 1:9 (10 %) → post-test 9:9 = 1:1 → probability 50 % – the same PPV, derived in one line. **Likelihood ratios** (LRs) are the portable form of the test; predictive values are local.

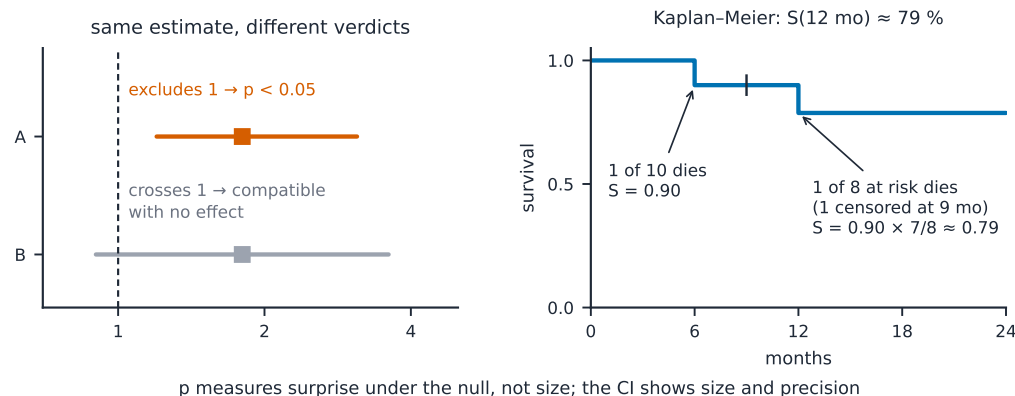
Common mistakes. Quoting sensitivity as “the test is 90 % right” (it is right about 90 % of the DISEASED – say of what). Applying a textbook PPV to a different-prevalence clinic. Confusing the two columns (sensitivity reads down the diseased column; specificity down the healthy column).

Summary. Se = $a/(a+c)$, Sp = $d/(b+d)$ – test properties; PPV/NPV – population properties, move with prevalence; natural-frequency trees make it visible; $LR+ = Se/(1-Sp)$ multiplies odds portably ($0.9/0.1 = 9$); 50 % PPV at 10 % prevalence despite Se = Sp = 90 %.

Statistical inference for clinical data

Why it matters. Inference is how a sample speaks about a population – and misread p-values and confidence intervals are the signature errors of the modern literature.

The idea, with a picture. A 95 % **confidence interval** is the range of true values compatible with the data at that level; a **p-value** P is the probability of data at least this extreme IF the null hypothesis were true ✓. The CI summarises size and precision; the p-value only summarises surprise.



The interval carries the estimate's size and precision; the survival curve steps down at events and drops at censoring gaps

The derivation (why the test follows the data type). Counts by group → χ^2 ; two means (normal-ish) → t-test (paired when each row is measured twice); several means → ANOVA; skewed/small-n → rank (non-parametric) alternatives; two continuous measurements → correlation (Pearson, or Spearman for ranks) ✓. The measurement scale decides the test – deciding after looking at results invites the wrong question.

Worked example 1 – reading one interval. (Fictional abstract) RR 1.8, 95 % CI 1.2–3.1: the interval excludes 1, so the data are incompatible with “no effect” at this level ($p < 0.05$ follows); the width says the estimate is imprecise but consistently above 1 ✓. An RR 1.8 with CI 0.9–3.6 says the opposite – same point estimate, different verdict, and only the interval tells you.

Worked example 2 – clinical link: the Kaplan–Meier arithmetic. (Synthetic) 10 patients enrolled; at 6 months 1 event among 10 at risk: $S = 9/10 = 0.90$; at 12 months 1 event among the 8 still at risk (one censored): $S = 0.90 \times 7/8 = 0.7875 \approx 79 \%$ ✓. Median survival is the time the curve crosses 50 %; censoring removes people from the denominator, not from the estimate's logic.

Common mistakes. “ $p > 0.05$ means no difference” (it means the data are compatible with no difference – absence of evidence $\neq$ evidence of absence). Reading the CI

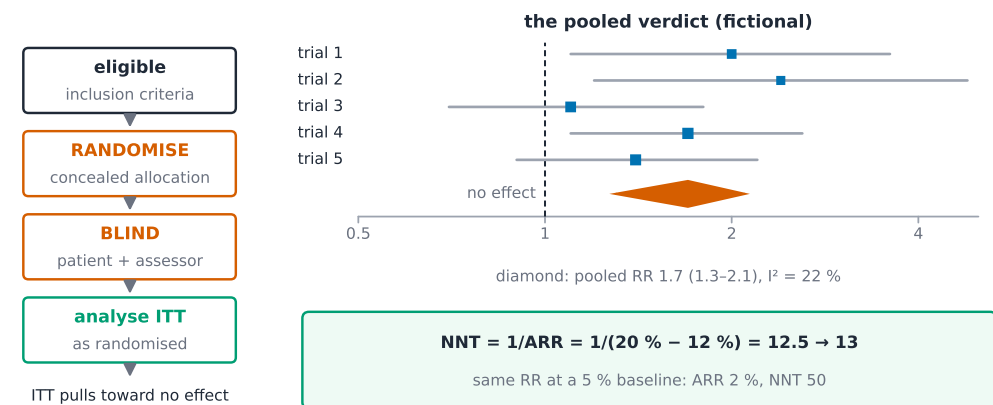
as a 95 % probability the truth lies inside (the interval is a property of the procedure, applied to this sample's estimate). Choosing the test after seeing which gives $p < 0.05$.

Summary. CI = compatible truths, carries size and precision; p-value = $P(\text{data this extreme} \mid \text{null})$; never “proven” or “no effect” from p alone; $\chi^2/t/\text{ANOVA}/\text{rank}/\text{correlation}$ chosen by data type; the **Kaplan–Meier curve** multiplies $(1 - \text{events/at-risk})$; median = 50 % crossing.

Clinical trials and systematic reviews

Why it matters. Trials are the machinery of proof and reviews are its assembly – the vocabulary (randomisation, blinding, ITT, NNT) is what every abstract and every forest plot assumes you know.

The idea, with a picture. A trial's spine: eligibility → randomisation (assignment by chance, allocation concealed) → blinding (participant, clinician, outcome assessor) → analysis by **intention to treat** (analyse as randomised, everyone counted) ✓. Reviews pool into a **forest plot** : squares (size = weight), lines (CI), diamond (pooled), and **heterogeneity** I^2 ✓.



Randomise, blind, analyse as randomised – then pool: squares weighted by precision, the diamond the pooled verdict

The derivation (why NNT is the clinical size of an effect). NNT

= $1/ARR$: control event rate 20 %, treated 12 % → ARR 8 % → NNT = $1/0.08 = 12.5 \rightarrow 13$ people treated for one to benefit ✓. NNT converts an abstract ratio into a workload – and its denominator (ARR) shrinks as baseline risk falls, which is why trial results do not transplant to lower-risk clinics unchanged.

Worked example 1 – the ITT logic. (Fictional trial) 20 % control vs 12 % treated: ARR 8 %, NNT 13 ✓. If 30 % of the treated arm stopped their assignment early, ITT still counts them as treated – diluting the apparent effect toward the null – because the question is the POLICY of assignment, not the biology of the adherent ✓. Per-protocol answers a different question and inherits adherence bias.

Worked example 2 – clinical link: reading the diamond. (Fictional forest) five trials, individual RRs 1.1–2.4 with wide CIs, pooled RR 1.7 (1.3–2.1), $I^2 = 22\%$ ✓: the diamond sits right of 1 and its interval excludes 1, with low heterogeneity – a consistent effect across settings. The same diamond with $I^2 = 78\%$ would prompt “why do these trials disagree?” before any pooling is trusted.

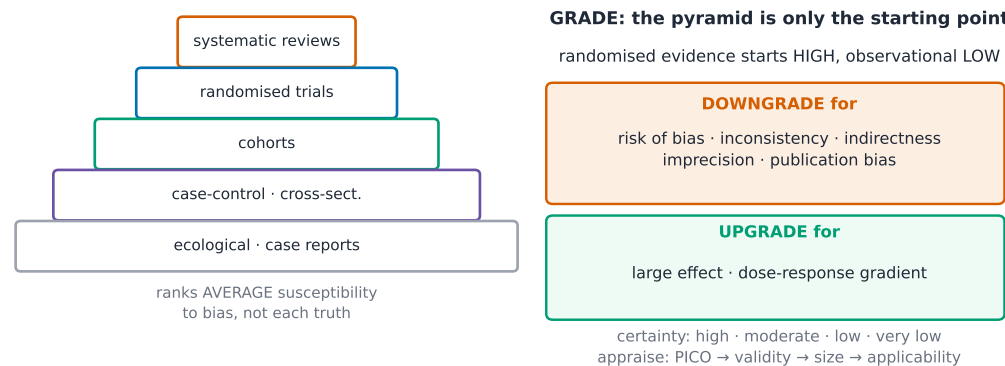
Common mistakes. Equating randomisation with blinding (chance balances groups; blindness balances expectations – a trial can have either alone). Reading per-protocol as the honest analysis (it breaks the randomisation’s protection). Treating the pooled diamond as gospel despite high heterogeneity.

Summary. Randomise (concealed) → blind → analyse as randomised (ITT, dilutes toward null); NNT = 1/ARR (20 % → 12 %: ARR 8 %, NNT 13); forest: squares weighted, diamond pooled, I^2 disagreement; low-baseline clinics shrink ARR and inflate NNT.

Critical appraisal and applying evidence

Why it matters. Evidence becomes medicine only at the bedside: appraised for validity, sized for importance, and matched to the patient in front of you. This unit is the course’s synthesis.

The idea, with a picture. The appraisal spine: **PICO** (population, intervention, comparison, outcome) framing → validity (were the results biased?) → size (how large, how precise?) → applicability (does my patient resemble the studied population?) ✓. Certainty grading (**GRADE**): randomised evidence starts high and observational evidence low; downgrade for risk of bias, inconsistency, indirectness, imprecision or publication bias; upgrade for large effects or a dose-response gradient ✓.



The pyramid ranks average susceptibility to bias, not the truth of each study – a flawed trial can mislead more than a good cohort

The derivation (why the hierarchy is a prior, not a verdict). Designs differ in their TYPICAL bias control: randomisation balances confounders measured and unmeasured; cohorts carry confounding; case-control adds recall bias ✓. But the hierarchy grades the design class, while the verdict belongs to the study in hand – a small, unblinded, per-protocol-analysed trial may deserve less trust than a large, pre-registered cohort with propensity adjustment. GRADE exists precisely because design class is only the starting point.

Worked example 1 – a fictional abstract in five questions. (Fictional) “Randomised, double-blind, n = 400; RR 0.75 (0.60–0.94)”: validity – randomised and blinded ✓; size – 25 % relative reduction, interval excludes 1; local applicability – were patients like mine enrolled (age, severity, co-medications)?; outcomes – patient-relevant or surrogate?; certainty – any downgrade features? Five questions, one structured verdict ✓.

Worked example 2 – clinical link: transplanting the NNT. (Fictional) a trial in severe disease (baseline 20 %, ARR 8 %, NNT 13); my patient’s clinic runs at 5 % baseline: with the same relative risk reduction the ARR scales to 8 % × 5/20 = 2 % → NNT 50 ✓. The relative effect may travel; the absolute burden – the workload and the harm-benefit arithmetic – is local.

Common mistakes. Appraising only for validity (a valid but tiny effect is still useless: size and precision decide importance). Grading by journal or design label alone (GRADE starts there and moves with the evidence’s flaws). Applying group averages to the patient whose defining feature was an exclusion criterion.

Summary. PICO → validity → size → applicability; GRADE: start from design, downgrade for bias/inconsistency/indirectness/imprecision, upgrade for large effects; the pyramid is a prior, not a verdict; relative effects travel, absolute benefits (ARR, NNT) are local – recompute before applying.