The Ultimate AKT Statistics Guide

for GP Trainees

From P-Values to Practice

A comprehensive AKT/MSRA revision guide covering every statistics topic you need to pass the exam — explained simply.

GP Core Revision

gpcorerevision.co.uk

Contents

How to Use This Guide

Welcome to the foundation. In the AKT, you don't need to be a mathematician; you need to be a BS detector. This guide combines high-yield explanations with worked examples to help you *feel* the numbers rather than memorise formulas.

Work through each chapter in order. The **Common AKT Traps** boxes highlight the exam's favourite pitfalls, and the **AKT-style questions** let you test your knowledge as you go. When you reach the end, challenge yourself with the **Final AKT Stats Mini-Quiz** and review the **What to Memorise for the AKT** summary.

This guide is best used actively — pause to predict answers before revealing them, sketch out 2×2 tables on paper, and talk through the analogies in your own words.

Part 1: Foundations

Chapter 1: P-Values

The "Fluke" Factor

The **P-value** stands for **Probability**. It measures the likelihood that a result happened purely by chance, *assuming there is no real effect*.

The Pub Explanation

Imagine playing heads-or-tails. If your mate wins ten times in a row, the probability of that happening by luck is tiny ($P < 0.05$). At that point you stop assuming "lucky streak" and start assuming "weighted coin".

A small P-value doesn't prove the effect is real — it just means *chance alone is a bad explanation*.

The Courtroom Analogy: Type 1 vs Type 2 Errors

Type 1 Error (False Positive) — *"Convicting an innocent man"* You conclude there IS a difference (significance), but actually there is NONE. *Clinical example: telling a healthy patient they have cancer.*

Type 2 Error (False Negative) — *"Letting a guilty man go free"* You conclude there is NO difference, but actually there IS one — you missed it. Usually caused by **insufficient power**, most commonly small sample size.

Power

Power is the probability that a study correctly detects a real effect (i.e. avoids a Type 2 error). Conventionally power is set at 80%. Power is driven by:

- **Sample size** (the big one)
- **Effect size** (bigger effects are easier to detect)
- **Variability** in the outcome
- **Significance threshold** (α, usually 0.05)

AKT-Style Question

A clinical trial compares Drug A with placebo for reducing migraine frequency. The study reports **P = 0.03**.

Which statement best describes this result?

A. There is a 3% chance the drug works B. There is a 97% chance the drug works C. The result is unlikely to be due to chance D. The drug is clinically effective

Correct answer: C

A P-value of 0.03 means that *if the drug had no real effect*, the probability of seeing a result this extreme (or more extreme) by chance alone is 3%. It does not tell you the

probability the drug works — only how surprising the result is under the assumption of no effect.

Common AKT Traps

- **Thinking P-value = probability the hypothesis is true.** It doesn't. It's the probability of the data assuming no effect.
- **Assuming "statistically significant" = "clinically important".** A tiny difference can be statistically significant in a huge study.
- **Believing P < 0.05 proves the treatment works.** It only suggests the result is unlikely due to chance.
- **Ignoring sample size.** Small studies often fail to reach significance even when a real effect exists (Type 2 error).
- **Misinterpreting P > 0.05.** It does not mean "no effect" — it means "not enough evidence to detect one".

You should now be able to:

- Explain what a P-value actually represents
- Recognise common misinterpretations
- Interpret P < 0.05 and P > 0.05 correctly
- Distinguish Type 1 from Type 2 errors
- Explain why statistical significance ≠ clinical significance

---

Chapter 2: Confidence Intervals

P-values give you a simple yes/no. **Confidence intervals (CIs)** give you the **magnitude** — and, crucially for the AKT, they let you spot the **line of no effect**.

Reading a 95% CI

A 95% confidence interval is the range of values within which we can be 95% confident the true effect lies.

- **Narrower CI** = more precision (usually larger sample size).
- **Wider CI** = less precision (usually smaller sample size).

The "Line of No Effect"

This is the killer concept for the AKT.

- **For ratios** (Relative Risk, Odds Ratio, Hazard Ratio): the line of no effect is **1**. → *"One is Done."*
- **For differences** (mean difference, ARR): the line of no effect is **0**. → *"Zero for Hero."*

If the 95% CI **crosses the line of no effect**, the result is **not statistically significant**, regardless of how impressive the point estimate looks.

Example

- RR = 0.7, 95% CI 0.5–0.9 → significant (does not cross 1)
- RR = 0.7, 95% CI 0.4–1.2 → **not** significant (crosses 1)
- Mean difference = 5 mmHg, 95% CI 2–8 → significant (does not cross 0)
- Mean difference = 5 mmHg, 95% CI −2 to 12 → **not** significant (crosses 0)

AKT-Style Question

A trial reports a relative risk of 0.78 with a 95% confidence interval of 0.65–0.94.

Which statement is most accurate?

A. The result is not statistically significant B. The result is statistically significant C. The CI is too wide to interpret D. The treatment increases risk

Correct answer: B

The CI does not cross 1, so the result is statistically significant. The RR of 0.78 suggests a 22% reduction in risk.

Common AKT Traps

- **Forgetting the line of no effect changes** depending on whether you're dealing with a ratio or a difference.
- **Treating a CI that just touches 1 (or 0) as significant.** It isn't — touching the line means the result is on the cusp.
- **Confusing "narrow CI" with "large effect".** Width tells you about precision, not size.
- **Ignoring the CI in favour of the P-value.** The CI tells you both significance *and* magnitude.

You should now be able to:

- Interpret a 95% CI for ratios and differences
- Identify when a CI indicates statistical significance
- Explain why narrower CIs = more precision
- Spot when a CI crosses the line of no effect

Chapter 3: Study Design

The AKT loves asking *"what type of study is this?"* — get the design right and you're halfway to the right answer about its strengths, weaknesses, and the kind of evidence it can produce.

The Four Big Designs

1. Randomised Controlled Trial (RCT) — *the gold standard* Patients are randomly allocated to intervention or control. Randomisation balances confounders, blinding re-

duces bias. Best for establishing **causation**. - *Pros:* Strongest evidence for cause and effect. - *Cons:* Expensive, time-consuming, sometimes unethical, may exclude real-world patients.

2. Cohort Study — *follow them forwards* Take a group of people, classify them by exposure (e.g. smokers vs non-smokers), and follow them over time to see who develops the outcome. - *Pros:* Good for rare exposures, gives incidence and relative risk. - *Cons:* Expensive, long follow-up, vulnerable to loss-to-follow-up.

3. Case-Control Study — *look backwards* Take a group with the disease (cases) and a group without (controls), then look back to compare exposures. - *Pros:* Quick and cheap; good for **rare diseases**. - *Cons:* Vulnerable to recall bias; gives odds ratios, not relative risk.

4. Cross-Sectional Study — *a snapshot* Measure exposure and outcome at the same time in a defined population. - *Pros:* Quick; good for measuring **prevalence**. - *Cons:* Can't establish cause and effect (chicken-and-egg problem).

The Hierarchy of Evidence

From strongest to weakest:

1. Systematic review / meta-analysis of RCTs
2. RCT
3. Cohort study
4. Case-control study
5. Cross-sectional study
6. Case series / case report
7. Expert opinion

Intention-to-Treat (ITT) vs Per-Protocol

Crops up in nearly every RCT question.

Intention-to-treat (ITT): Analyse patients in the group they were originally allocated to, *regardless of whether they actually took the treatment*. Preserves randomisation. Gives a more conservative, real-world estimate. **This is the preferred analysis for RCTs.**

Per-protocol: Analyse only those who completed the treatment as planned. Can exaggerate treatment effect because non-compliers are removed.

> ***Memory hook:*** *ITT = "what happens when you* prescribe *it" (pragmatic). Per-protocol = "what happens when you* take *it" (explanatory).*

AKT-Style Question

A new treatment for resistant hypertension is being investigated. The condition is rare, and follow-up over several years is impractical. Which study design is most appropriate?

A. Randomised controlled trial B. Prospective cohort study C. Case-control study D. Cross-sectional study

Correct answer: C

For rare diseases where prospective follow-up is impractical, case-control studies are the most efficient design. You identify cases retrospectively and compare exposures to matched controls.

Common AKT Traps

- **Confusing cohort and case-control.** Cohort = exposure → outcome (forwards). Case-control = outcome → exposure (backwards).
- **Calling a cross-sectional study an "incidence study".** Cross-sectional measures *prevalence*, not incidence.
- **Assuming RCTs are always best.** They aren't always feasible or ethical — and the wrong RCT can be worse than a good observational study.
- **Mixing up ITT and per-protocol.** ITT is the conservative, randomisation-preserving choice.
- **Forgetting case-control studies give odds ratios, not relative risk.**

You should now be able to:

- Identify the four major study designs from a stem
- Recognise the hierarchy of evidence
- Match study design to research question
- Distinguish ITT from per-protocol analysis

Chapter 4: Bias & Validity

The AKT is full of questions that essentially ask: *"What's wrong with this study?"* Knowing the named biases and what they look like is high-yield revision.

The Main Types of Bias

Selection bias — when the people in the study aren't representative of the population you want to generalise to. *Example:* Recruiting only hospital inpatients for a study about a community-level intervention.

Recall bias — when participants' memory of past exposures is influenced by whether they have the outcome. *Example:* Mothers of children with congenital abnormalities recalling pregnancy exposures more thoroughly than mothers of healthy children. Classic in case-control studies.

Measurement (information / observer) bias — when the way outcomes or exposures are measured differs between groups. *Example:* The investigator knows who got the drug and assesses outcomes more favourably (avoided by blinding).

Attrition bias — when participants who drop out differ systematically from those who stay. *Example:* In an antidepressant trial, patients who feel worse drop out, leaving only those who tolerated the drug well.

Publication bias — positive studies are more likely to be published than negative ones (covered in Chapter 12).

Lead-time bias — a screening test detects disease earlier, making survival *look* longer even if death occurs at the same time. Classic in screening questions.

Length-time bias — screening preferentially detects slow-growing, less aggressive disease (because fast-growing tumours appear and kill between screens). Makes screening look more effective than it is.

Validity & Reliability

Internal validity — does the study correctly answer its own question? (Free from bias and confounding within the study.)

External validity (generalisability) — do the results apply to *other* populations beyond the study?

Reliability — does the measurement give the *same* result on repeat testing? (Consistency.)

Validity — does the measurement actually measure what it claims to? (Accuracy.)

> ***The dartboard analogy.*** *Reliable = arrows all in the same spot. Valid = arrows all on the bullseye. You can be reliable without being valid (consistently wrong).*

Reducing Bias

- **Randomisation** — balances confounders between groups.
- **Allocation concealment** — recruiters don't know which group the next patient will be assigned to.
- **Blinding** — single (patient), double (patient + investigator), triple (patient + investigator + analyst).
- **ITT analysis** — preserves randomisation.
- **Standardised outcome assessment** — same tools, same training, same blinding.

AKT-Style Question

A case-control study investigates whether anti-emetic use in pregnancy is associated with cleft palate. Mothers of affected babies recall their exposures more thoroughly than mothers of unaffected babies. Which bias is most likely?

A. Selection bias B. Recall bias C. Lead-time bias D. Attrition bias

Correct answer: B

Recall bias is the classic problem in case-control studies — mothers of affected children are likely to scrutinise their pregnancy exposures far more than control mothers,

exaggerating associations.

Common AKT Traps

- **Mixing up validity and reliability.** Valid = right. Reliable = consistent.
- **Confusing internal and external validity.** Internal = within the study. External = generalisability.
- **Forgetting lead-time and length-time bias** — they only really apply to screening but they come up.
- **Assuming blinding fixes everything.** It reduces measurement bias but doesn't address selection bias.
- **Treating "randomised" and "blinded" as synonyms.** Randomisation deals with confounders at the start; blinding deals with bias during follow-up.

You should now be able to:

- Name and recognise the main types of bias
- Distinguish validity from reliability
- Distinguish internal from external validity
- Explain how randomisation, blinding, and ITT reduce bias

Part 2: Diagnostic Tests

Chapter 5: The 2×2 Table & Diagnostic Test Properties

The Fishing Net Analogy

A diagnostic test is a **fishing net**.

Sensitivity (the net): How well does the net catch all the fish? High sensitivity = catches all the true positives (but also some junk). **Rules OUT disease (SnOut)**.

Specificity (the sorter): How well does the net throw the rubbish back? High specificity = correctly identifies true negatives. **Rules IN disease (SpIn)**.

The 2×2 Table: Coeliac Disease Example

You test 1,000 patients in a hospital clinic where coeliac prevalence is 10%.

	Disease Present (Sick)	Disease Absent (Healthy)	Totals
Test Positive (+)	90 (True Positive)	45 (False Positive)	135
Test Negative (–)	10 (False Negative)	855 (True Negative)	865
Totals	100	900	1,000

From this table:

- **Sensitivity** = TP / (TP + FN) = 90 / 100 = **90%**

- **Specificity** = TN / (TN + FP) = 855 / 900 = **95%**
- **Positive Predictive Value (PPV)** = TP / (TP + FP) = 90 / 135 = **67%**
- **Negative Predictive Value (NPV)** = TN / (TN + FN) = 855 / 865 = **99%**

Memory Aid

> ***SnNOut**: a highly **SeN**sitive test, when **N**egative, rules **OUT** disease. **SpPIn**: a highly **SP**ecific test, when **P**ositive, rules **IN** disease.*

Common AKT Traps

- **Confusing sensitivity with PPV.** Sensitivity tells you how well the test detects disease — not how likely a positive result is to be true.
- **Forgetting which way the calculations go.** Always draw the 2×2 with disease across the top and test result down the side, and the numbers will fall out.
- **Treating sensitivity and PPV as interchangeable.** They aren't — PPV depends on prevalence (see next chapter).
- **Assuming high sensitivity = rules in disease.** It's the opposite (SnOut).

You should now be able to:

- Calculate sensitivity, specificity, PPV and NPV from a 2×2 table
- Apply SnOut and SpIn correctly
- Interpret 2×2 tables confidently

Chapter 6: The Prevalence Trap

Why good tests fail in primary care

Sensitivity and specificity are fixed properties of the test. PPV and NPV are not — they change with prevalence.

This is the single most important diagnostic-stats concept for GP practice, and the AKT loves testing it.

Worked Example: Same Test, Different Settings

A test for coeliac disease has **sensitivity 90%** and **specificity 95%**. Look what happens in different settings:

Hospital clinic (prevalence 10%, 1,000 patients):

	Disease +	Disease –	Total
Test +	90	45	135
Test –	10	855	865
PPV = 90/135 = 67%			

GP population (prevalence 1%, 1,000 patients):

	Disease +	Disease –	Total
Test +	9	49.5	58.5
Test –	1	940.5	941.5
PPV = 9/58.5 = 15%			

Same test. In hospital, two-thirds of positives are real. In GP, **only one in seven**. Most of your positive results will be false positives.

Why this matters clinically

- Don't trust a positive test result on its own when prevalence is low.
- This is why hospital tests can perform poorly when imported into GP without thought.
- It's also why **pre-test probability** (what you thought before the test) matters so much — covered in the next chapter.

AKT-Style Question

A new test for coeliac disease has a sensitivity of 90% and a specificity of 95%. In a GP population where the prevalence of coeliac disease is 1%, a patient tests positive.

Which statement is most accurate?

A. The patient almost certainly has coeliac disease B. The test has ruled the disease in C. Most positive results will still be false positives D. The test is not sensitive enough for primary care

Correct answer: C

Even with excellent sensitivity and specificity, a low-prevalence setting like GP massively reduces the PPV. When prevalence is only 1%, most positive results will be false positives — the classic Prevalence Trap.

Common AKT Traps

- **Ignoring prevalence.** PPV and NPV change dramatically with prevalence. Sensitivity and specificity do not.
- **Assuming a "good test" works everywhere.** A hospital-grade test can perform poorly in GP because of low prevalence.
- **Forgetting false positives dominate in low-prevalence settings.** Even a tiny false-positive rate overwhelms true positives when disease is rare.

You should now be able to:

- Explain how prevalence affects PPV/NPV
- Predict whether a test will perform well or badly in GP
- Justify why hospital-grade tests sometimes fail in primary care

---

Chapter 7: Likelihood Ratios & The Fagan Nomogram

What an LR actually tells you

A **likelihood ratio (LR)** is the "shove" that moves you from **pre-test probability** to **post-test probability**.

- **LR+** pushes you *towards* the diagnosis (how much a positive result increases the probability of disease).
- **LR–** pushes you *away from* the diagnosis (how much a negative result decreases it).

The Golden Rules

Value	Interpretation
LR+ > 10	Strong rule-in
LR+ 5–10	Moderate rule-in
LR+ 2–5	Weak rule-in
LR ≈ 1	Useless
LR– 0.2–0.5	Weak rule-out
LR– 0.1–0.2	Moderate rule-out
LR– < 0.1	Strong rule-out

Formulas (for understanding, not memorisation)

> *LR+ = Sensitivity / (1 – Specificity) LR– = (1 – Sensitivity) / Specificity*

Clinical Example

D-dimer has **sensitivity 95%**, **specificity 50%**.

- LR+ = 0.95 / 0.50 = **1.9** (weak — useless for ruling PE in).
- LR– = 0.05 / 0.50 = **0.1** (strong — excellent for ruling PE out).

This is exactly why D-dimer is used as a **rule-out test** when pre-test probability is low.

Pre-test → Post-test Probability: The Fagan Nomogram

The Fagan nomogram is a graphical tool that lets you combine three things to get a clinical answer:

1. **Pre-test probability** (your clinical suspicion before the test, often estimated from prevalence or clinical scoring like Wells' score)
2. **Likelihood ratio** (from the test)
3. **Post-test probability** (the answer)

Draw a line from your pre-test probability, through the LR, and read off the post-test probability.

Why this matters for the AKT and for real life

LRs are independent of prevalence, but **post-test probability is not** — it depends entirely on what you thought before the test. The same positive test result means very different things in a 70-year-old smoker with chest pain (high pre-test probability) versus a 25-year-old with the same symptom (low pre-test probability).

Rule of thumb without the nomogram:

- LR+ of 10 increases probability by about **45 percentage points**
- LR+ of 5 increases probability by about **30 percentage points**
- LR+ of 2 increases probability by about **15 percentage points**
- LR– of 0.1 decreases probability by about **45 percentage points**
- LR– of 0.5 decreases probability by about **15 percentage points**

AKT-Style Question

A test for pulmonary embolism has a sensitivity of 95% and a specificity of 50%. What is the most appropriate interpretation of its likelihood ratios?

A. LR+ is high, so the test is good for ruling in PE B. LR– is high, so the test is poor for ruling out PE C. LR+ is low, so a positive result does not increase the probability of PE much D. LR– is high, so a negative result increases the probability of PE

Correct answer: C

With sensitivity 95% and specificity 50%, the LR+ is around 1.9 — too low to rule in disease. The LR– is around 0.1, strong for ruling out. This is why D-dimer is an excellent rule-out tool but a poor rule-in.

Common AKT Traps

- **Thinking LR+ depends on prevalence.** It doesn't. LR+ and LR– are properties of the test.
- **Confusing LR+ with PPV.** LR+ tells you how much a positive result shifts probability; PPV tells you how likely the patient actually has the disease.
- **Assuming a high-sensitivity test rules in disease.** High sensitivity rules out (SnOut). High specificity rules in (SpIn).
- **Forgetting the magic numbers:** LR+ > 10 = strong rule-in; LR– < 0.1 = strong rule-out; LR ≈ 1 = useless.

You should now be able to:

- Interpret LR+ and LR– values
- Explain why LRs are properties of the test, not the population
- Combine pre-test probability with an LR to estimate post-test probability
- Recognise why the same test result means different things in different patients

---

Chapter 8: Screening Principles

Screening is testing **asymptomatic** people for a disease. It's a deceptively complex topic because the benefits are easy to overstate and the harms are easy to overlook. The AKT loves screening questions, particularly around the Wilson & Jungner criteria and screening-specific biases.

Wilson & Jungner Criteria (1968)

The classic ten principles for whether a screening programme is worthwhile. Memorise the headlines:

About the disease: 1. The condition should be an **important health problem**. 2. There should be a **recognisable latent or early symptomatic stage**. 3. The **natural history** of the condition should be adequately understood.

About the test: 4. There should be a **suitable test** (sensitive, specific, acceptable). 5. The test should be **acceptable** to the population.

About the treatment: 6. There should be an **accepted treatment** for patients with the disease. 7. Facilities for **diagnosis and treatment** should be available. 8. There should be an **agreed policy** on whom to treat.

About the programme: 9. The **cost** should be economically balanced against expenditure on care. 10. **Case-finding** should be a continuing process, not a one-off.

> ***Memory hook:*** *disease, test, treatment, programme — four buckets, two or three points each.*

Screening-Specific Biases

Lead-time bias — Screening detects disease earlier. Survival from diagnosis *looks* longer, even if the time of death is unchanged.

Length-time bias — Screening preferentially detects slow-growing, less aggressive disease. Aggressive cancers tend to appear and kill between screening rounds, so the cases caught look more favourable than the underlying disease really is.

Selection bias (the "healthy screenee" effect) — People who attend screening tend to be healthier overall than those who don't. Outcomes look better partly because of *who* turns up, not what the test does.

Overdiagnosis — Detecting disease that would never have caused harm in the patient's lifetime. Leads to overtreatment of indolent conditions (a major issue in PSA screening, for example).

The UK National Screening Committee Principles

The UK NSC has expanded Wilson & Jungner to include more rigorous evidence requirements — particularly around RCT evidence that screening reduces mortality and avoids overdiagnosis. For the AKT, the original Wilson & Jungner list is what you're most likely to be tested on.

AKT-Style Question

A new screening test for an indolent cancer is shown to "improve five-year survival from 60% to 85%" when introduced. There is no change in overall mortality.

Which bias most likely explains this finding?

A. Recall bias B. Lead-time bias C. Selection bias D. Attrition bias

Correct answer: B

If overall mortality is unchanged but five-year survival from diagnosis improves, you're seeing lead-time bias: the test diagnoses disease earlier, the clock starts ticking earlier, and survival *from diagnosis* gets longer — but patients still die at the same age.

Common AKT Traps

- **Confusing lead-time and length-time bias.** Lead-time = earlier diagnosis, same death. Length-time = preferentially catching slow-growers.
- **Forgetting that "improved survival" doesn't mean "lower mortality".** They're different endpoints. The AKT often tests this distinction.
- **Mixing up screening and case-finding.** Screening targets asymptomatic populations; case-finding looks at people with relevant risk factors or symptoms.
- **Assuming all positive screening tests need treatment.** Overdiagnosis means some shouldn't.

You should now be able to:

- Recall the headline Wilson & Jungner criteria
- Distinguish lead-time, length-time, and selection bias in screening
- Explain overdiagnosis and why it matters
- Recognise why "improved survival" alone is a weak endpoint for screening

---

Part 3: Treatment Effects

Chapter 9: Risk & Reward — ARR, RRR, NNT & NNH

Drug companies love to quote dramatic-sounding **relative risk reductions (RRR)**. Clinicians need the more honest measures: **absolute risk reduction (ARR)** and **number needed to treat (NNT)**. The AKT will test whether you can see through marketing spin.

The Marketing Trick: RRR vs ARR

Imagine a drug that cuts heart attack risk from 2% to 1% over 5 years.

- **Relative Risk Reduction (RRR)** = (2 – 1) / 2 = **50%** ("Cuts your risk in half!")
- **Absolute Risk Reduction (ARR)** = 2 – 1 = **1 percentage point** ("99 out of 100 people gained nothing.")

Same data. Wildly different message. The AKT will test whether you can spot when RRR is being used to inflate a small effect.

Number Needed to Treat (NNT)

> ***NNT = 100 / ARR (when ARR is expressed as a percentage)*** *Equivalent: NNT = 1 / ARR (when ARR is expressed as a decimal)*

NNT answers: *"How many people do I have to treat for one person to benefit?"*

- ARR of 2% → NNT = 100 / 2 = **50**
- ARR of 1% → NNT = 100 / 1 = **100**
- ARR of 25% → NNT = 100 / 25 = **4**

Lower NNT = more impressive treatment. Statins for primary prevention have an NNT of around 100 over 5 years. Antibiotics for sore throat have an NNT in the thousands.

Number Needed to Harm (NNH)

The mirror image of NNT.

> ***NNH = 100 / ARI*** *(Absolute Risk Increase of an adverse event, expressed as a percentage)*

NNH answers: *"How many people do I have to treat for one to be harmed?"*

A treatment is only worthwhile if **NNT < NNH** for the harm you're worried about. This is the heart of shared decision-making.

Worked example: A new anticoagulant prevents stroke in 4% of patients over 5 years (ARR = 4%) but causes major bleeding in 1% (ARI = 1%).

- NNT = 100 / 4 = **25** (treat 25 to prevent one stroke)
- NNH = 100 / 1 = **100** (treat 100 to cause one major bleed)

Favourable trade-off: NNT < NNH.

AKT-Style Question

A statin reduces the risk of cardiovascular events from 4% to 2% over 5 years. The drug company advertises this as a "50% reduction in risk".

Which statement best describes this claim?

A. It is accurate and reflects the absolute risk reduction B. It is misleading because it uses relative risk reduction C. It is incorrect because the NNT is 2 D. It is incorrect because the absolute risk reduction is 50%

Correct answer: B

The risk drops from 4% to 2% — an absolute risk reduction (ARR) of 2 percentage points. The company quotes the relative risk reduction (RRR), which is 50%. RRR often sounds dramatic, but ARR and NNT give a more realistic picture of benefit (NNT here = 50).

Common AKT Traps

- **Confusing RRR with ARR.** RRR often looks impressive. ARR tells you the real-world benefit.
- **Calculation slip on NNT.** NNT = 100 / ARR (in %). ARR of 2% → NNT = 50, not 100.
- **Assuming a large RRR means a large clinical effect.** A 50% reduction from 0.2% to 0.1% is tiny in absolute terms.
- **Forgetting NNH.** A treatment is only worthwhile if the NNT is meaningfully lower than the NNH for relevant harms.
- **Believing ARR and NNT are fixed properties of a treatment.** They vary with baseline risk — higher baseline risk → lower NNT (bigger benefit).

You should now be able to:

- Calculate ARR and RRR from raw event rates
- Calculate NNT from ARR
- Calculate NNH from an absolute risk increase of harm
- Compare NNT and NNH to judge whether treatment is worthwhile

---

Chapter 10: Hazard Ratios & Kaplan–Meier Curves

Kaplan–Meier Curves: Survival Over Time

Kaplan–Meier (K–M) curves are used when the outcome isn't just "did you die?" but **"how long until you die?"** They are perfect for cancer trials, cardiovascular studies, and any condition where timing matters.

What the curve shows:

- **Y-axis** = proportion surviving (1.0 at start, decreasing over time)
- **X-axis** = time
- Each **downward step** = an event (e.g. death)
- **Flatter curve** = better survival
- **Steeper curve** = worse survival

Censoring happens when we stop knowing what happened to a patient — they left the study, moved away, or were still alive when the study ended. Shown on the graph as **small tick marks**. They are *not* deaths.

Comparing Two Curves

- The curve that stays **higher** = better survival.
- Look for **consistent separation** between curves — that's strong evidence of a real difference.
- **Crossing curves** are a red flag: the proportional hazards assumption breaks down and the hazard ratio becomes unreliable.

Hazard Ratios (HR)

The hazard ratio is the **numerical summary** of what the K–M curves show visually. It compares the **rate of events over time** between two groups.

- **HR = 1** → no difference
- **HR < 1** → treatment reduces the event rate (good for survival outcomes)
- **HR > 1** → treatment increases the event rate (bad)

> ***Crucial distinction:*** *the hazard ratio is about* rate*, not total events at the end. HR of 0.7 means the* rate *of events is 30% lower, not necessarily that 30% fewer people died overall.*

HR vs Relative Risk

Both compare risk, but only HRs include **time**.

- **RR**: "By the end of the study, did you have the event?" (yes / no)
- **HR**: "How fast are events happening in Group A vs Group B?"

AKT-Style Question

A trial compares Drug A with placebo for improving survival in heart failure. The Kaplan–Meier curves show clear and consistent separation, with the Drug A curve remaining higher throughout. The hazard ratio is 0.70 with a 95% CI of 0.55–0.90.

Which statement best describes this result?

A. Drug A reduces the risk of death by 30% at the end of the study B. Drug A reduces the rate of death over time compared with placebo C. Drug A improves survival only at the final time point D. The result is not statistically significant

Correct answer: B

An HR of 0.70 means the *rate* of death over time is 30% lower with Drug A. Consistent K–M curve separation supports this. The CI (0.55–0.90) doesn't cross 1, so it's significant.

Common AKT Traps

- **Thinking the hazard ratio is the same as relative risk.** HR reflects event rate over time, not the final proportion.
- **Ignoring curve separation.** Consistent separation = strong evidence. Crossing curves = unreliable HR.
- **Misinterpreting censoring.** Tick marks are not deaths — they mean follow-up ended.
- **Assuming HR < 1 always means a big effect.** It means benefit, but the CI tells you precision and significance.

You should now be able to:

- Read a Kaplan–Meier curve and interpret censoring
- Interpret HR < 1 and HR > 1

- Recognise when HR is unreliable (crossing curves)
- Distinguish HR from relative risk

Part 4: Advanced Topics

Chapter 11: Regression & Odds Ratios

Regression models help us understand how different factors affect an outcome. For the AKT, you don't need to calculate anything — you just need to **interpret** the numbers.

Odds Ratios (OR)

An odds ratio compares the **odds** of an outcome between two groups.

- **OR = 1** → no difference
- **OR > 1** → higher odds of the outcome
- **OR < 1** → lower odds of the outcome

Examples: - OR = 2.0 → "Twice the odds of the outcome." - OR = 0.5 → "Half the odds of the outcome."

Important: **odds ≠ risk**. When the outcome is common, OR exaggerates risk. When it's rare, OR ≈ RR.

Adjusted vs Unadjusted Models

- **Unadjusted model**: looks at one factor at a time.
- **Adjusted model**: accounts for other variables that might influence the result (e.g. age, sex, smoking, BMI).

If the OR **changes a lot** after adjustment → **confounding** was present.

Confounding

A confounder is a variable that: - affects the exposure, **and** - affects the outcome, **and** - distorts the apparent relationship between them.

> ***Classic example:*** *coffee drinking appears linked to lung cancer — until you adjust for smoking. Smoking is the confounder.*

Reading a Regression Table

A typical table includes:

Variable	OR	95% CI	P-value
Age (per 10 years)	1.5	1.2–1.8	< 0.001
Current smoker	2.2	1.6–3.0	< 0.001
BMI (per 5 units)	1.3	1.0–1.6	0.04

Variable	OR	95% CI	P-value
Regular exercise	0.7	0.5–0.9	0.01

How to read it: - If the CI **does not cross 1** → statistically significant. - OR > 1 → increases odds; OR < 1 → decreases odds. - After adjusting for age, smoking, BMI, and exercise, each remains an independent predictor.

AKT-Style Question

A study looks at the association between smoking and chronic cough. The unadjusted OR is 2.8. After adjusting for age, the OR falls to 1.6.

What is the most likely explanation?

A. Age is a confounder of the relationship between smoking and chronic cough B. Smoking is not associated with chronic cough C. The adjusted model is incorrect D. The sample size is too small

Correct answer: A

The OR drops from 2.8 to 1.6 after adjusting for age, meaning part of the apparent effect of smoking was actually due to age. This is classic confounding: age is linked to both smoking and chronic cough, and adjusting for it reveals the true, smaller effect.

Common AKT Traps

- **Confusing correlation with causation.** Regression shows associations, not proof of cause.
- **Ignoring confounders.** A large change in OR after adjustment is a flag for confounding.
- **Misreading the CI.** A CI that crosses 1 means not significant — even if the OR looks impressive.
- **Treating OR as risk.** Odds and risk are only similar when the outcome is rare.
- **Over-interpreting tiny P-values.** A small P-value doesn't mean a big or important effect.

You should now be able to:

- Interpret odds ratios from logistic regression
- Explain why odds ≠ risk
- Recognise confounding by comparing unadjusted and adjusted ORs
- Identify which predictors are significant from a regression table

---

Chapter 12: Forest Plots & Meta-Analysis

Anatomy of a Forest Plot

Each **horizontal line** = one study (the line is its 95% CI, the box is its point estimate). The **size of the box** = the study's weight in the analysis (bigger = more weight). The **diamond at the bottom** = the **pooled effect** across all studies.

The **line of no effect** runs vertically through the plot: - At **1** for ratios (RR, OR, HR) - At **0** for differences (mean difference)

If the diamond **crosses** the line of no effect → not significant. If it sits entirely to one side → significant.

Heterogeneity (I^2): The Fruit Salad Test

I^2 measures how much the studies disagree.

I^2	Interpretation
< 25%	Low heterogeneity — "apples with apples", safe to pool
25–75%	Moderate heterogeneity — interpret with caution
> 75%	High heterogeneity — "fruit salad", pooling is unreliable

High I^2 means the studies are too different to combine meaningfully. You can still report the pooled estimate, but the AKT will expect you to flag it as unreliable.

Publication Bias and the Funnel Plot

Publication bias: positive (statistically significant) studies are more likely to be published than negative ones.

The **funnel plot** is the visual check. - **Symmetrical funnel** → no publication bias. - **Asymmetrical funnel** (corner missing, usually the bottom-left) → likely publication bias; the missing studies are negative ones that never got published.

AKT-Style Question

A meta-analysis evaluates whether a new antihypertensive reduces stroke risk. The pooled effect estimate (diamond) shows a relative risk of 0.85 with a 95% CI of 0.78–0.92. The diamond lies entirely to the left of the line of no effect.

Which statement best describes this result?

A. The treatment has no significant effect on stroke risk B. The treatment significantly reduces stroke risk C. The result is not reliable because the diamond crosses 1 D. The individual studies must all show benefit

Correct answer: B

The pooled estimate is entirely to the left of the line of no effect (1.0), and the CI does not cross 1. The treatment significantly reduces stroke risk. Individual studies may vary — the pooled estimate is what matters.

Common AKT Traps

- **Thinking every individual study must show benefit.** Meta-analysis pools results — some studies may favour placebo.
- **Misreading the diamond.** If it crosses the line of no effect, the pooled result is not significant.
- **Confusing study weight with effect size.** Bigger boxes = more weight, not bigger effect.
- **Ignoring heterogeneity.** High I^2 means studies disagree — pooled results should be interpreted cautiously.
- **Forgetting the line of no effect changes.** Ratios → 1. Differences → 0.
- **Treating "statistically significant" as "clinically important".** Not the same thing.

You should now be able to:

- Identify the pooled effect (diamond) in a forest plot
- Interpret whether the diamond crosses the line of no effect
- Explain study weights (box sizes)
- Recognise high heterogeneity (I^2) and its implications
- Spot publication bias on a funnel plot

Final Revision Section

What to Memorise for the AKT

1. **P-values:** $P < 0.05$ = statistically significant. P-value ≠ probability the hypothesis is true.
2. **Confidence intervals:** For ratios → CI crossing **1** = not significant. For differences → CI crossing **0** = not significant.
3. **SnOut / SpIn:** High **Se**n**N**sitivity → rules **OUT** disease. High **SP**ecificity → rules **IN** disease.
4. **Prevalence:** PPV and NPV change with prevalence. Sensitivity and specificity do not.
5. **Likelihood ratios:** LR+ > 10 = strong rule-in. LR– < 0.1 = strong rule-out. LR ≈ 1 = useless.
6. **Hazard ratios:** HR < 1 = reduced event rate over time. HR > 1 = increased event rate.
7. **Absolute vs relative risk:** ARR = real-world benefit. RRR often exaggerates.
8. **NNT:** NNT = 100 / ARR (when ARR is in percentage points). NNH is the mirror image.
9. **Forest plots:** Diamond = pooled effect. Diamond crossing line of no effect = not significant.
10. **Kaplan–Meier:** Higher curve = better survival. Crossing curves = unreliable HR.
11. **Study design:** RCT (gold standard); cohort (forwards); case-control (backwards, rare diseases); cross-sectional (snapshot, prevalence).

12. **Bias types:** selection, recall, measurement, attrition, lead-time, length-time, publication. Validity ≠ reliability.
13. **ITT vs per-protocol:** ITT preserves randomisation and is the default for RCTs.
14. **Wilson & Jungner:** screening criteria across disease, test, treatment, and programme.

Final AKT Stats Mini-Quiz

Time yourself: 12 questions, 12 minutes. Answers and explanations follow.

1. A study reports P = 0.049. Which interpretation is most accurate? A. There is a 4.9% chance the result is true B. There is a 95.1% chance the treatment works C. The result is unlikely to be due to chance D. The study proves the hypothesis

2. A 95% CI for a risk ratio is 0.92–1.01. What does this mean? A. The treatment is effective B. The result is not statistically significant C. The treatment increases risk D. The CI is too narrow to interpret

3. A test has sensitivity 98% and specificity 40%. In GP, prevalence is 1%. What is true? A. PPV will be high B. Most positives will be false positives C. The test is good for ruling in disease D. The test is useless

4. A test has LR+ = 2 and LR– = 0.05. Which is true? A. Good for ruling in B. Good for ruling out C. Useless for both D. LR+ and LR– must match in strength

5. In a Kaplan–Meier curve, the treatment curve stays above the control curve, but they cross briefly at 6 months. What does this imply? A. HR cannot be interpreted reliably B. The treatment is harmful C. The treatment is effective throughout D. Censoring is incorrect

6. A hazard ratio of 0.65 means: A. 35% fewer total events at study end B. 35% lower event rate over time C. The treatment is clinically important D. The treatment halves mortality

7. A drug reduces risk from 12% to 9%. What is the NNT? A. 11 B. 33 C. 3 D. 9

8. In a forest plot, the pooled diamond touches but does not cross the line of no effect. What does this mean? A. Significant B. Not significant C. Clinically important D. The studies are biased

9. A logistic regression shows an odds ratio of 1.8 for smoking and chronic cough. What does this mean? A. Smokers have 80% higher odds of cough B. Smokers have 80% higher risk of cough C. Smoking causes cough D. The model is linear

10. A meta-analysis shows high heterogeneity (I^2 = 78%). What is the correct interpretation? A. The pooled result is highly reliable B. The studies are too different to combine confidently C. Publication bias is present D. The diamond must cross the line of no effect

11. A new screening test for a slow-growing cancer improves five-year survival from 60% to 85%, but overall mortality is unchanged. Which bias most likely explains this? A. Selection bias B. Recall bias C. Lead-time bias D. Attrition bias

12. A case-control study investigates a possible link between an antibiotic and tendon rupture. Which design feature most reduces recall bias? A. Larger sample size B. Blinding of outcome assessors C. Using prescription records rather than patient interviews to establish exposure D. Adjusting for confounders in the analysis

---

Mini-Quiz Answers

1. C. A P-value of 0.049 means that under the assumption of no effect, the probability of seeing data this extreme is 4.9%. It doesn't tell you the probability the hypothesis is true.

2. B. The CI (0.92–1.01) crosses 1, so the result is not statistically significant — even though it's close.

3. B. With prevalence of just 1% and specificity of 40%, the false-positive rate dominates. Most positives will be false positives — the Prevalence Trap.

4. B. LR+ of 2 is too weak to rule in. LR– of 0.05 is excellent for ruling out (well under 0.1).

5. A. Crossing curves break the proportional hazards assumption — the HR becomes unreliable.

6. B. HR reflects the rate of events over time, not total events at study end. A 35% lower rate is what 0.65 means; whether this is "clinically important" depends on context.

7. B. ARR = 12% – 9% = 3%. NNT = 100 / 3 ≈ 33.

8. B. Touching is not crossing, but in standard convention a CI that includes the line of no effect (even just touching) is treated as not statistically significant.

9. A. OR of 1.8 means 80% higher *odds*, not 80% higher *risk*. (When the outcome is rare, the two are similar, but the AKT will test the distinction.)

10. B. $I^2 > 75\%$ means high heterogeneity — the studies are too different to combine confidently.

11. C. Lead-time bias: earlier detection lengthens survival from diagnosis without changing the time of death. Overall mortality is the giveaway.

12. C. Recall bias is reduced when exposure data come from objective records rather than patient memory. Prescription records are the cleanest source.

GP Core Revision Series | Medical Statistics