Statistics for the GP trainee

1. Statistics for the GP trainee

As improbable as it seems, we were told there was a need for an article on medical statistics – like a cat asking for swimming lessons!

If you don't know your median from your mode, and your confidence interval has reached an all-time low, this article is just the thing for you.

Reading this article *might* make your statistics knowledge better, but let's not jump to conclusions: it could just be an association. If you want to know why, read on.

This article was updated in May 2025 and now includes the previous article, More statistics for the GP trainee.

1.1. Averages (mean, median and mode)

Mean (often called '*the average*'): add all the results up and divide by the number of results. Used in the calculation of 'average home BP'. The key disadvantage of this method is that an outlier can significantly skew the mean. To avoid this, we use median which reduces the impact of outliers.

Median: line up all the numbers in order. The median is the middle one. Outliers do not unduly influence the result.

Mode: the most frequently occurring value in a series. Useful measure for nominal data (things with names, e.g. apples and oranges or hair colour). Not used that often in academic papers.

1.2. Probability (p-value)

I have a coin which, when tossed, always lands on heads. I toss it once. Heads. Does that prove my hypothesis?

This illustrates the p-value and the concept of significance. The p-value (probability) here is 50:50 (or p=0.5). Clearly, this is not significant, and won't convince anyone that my coin always lands on heads.

I toss it again: heads. This still could be lucky chance. I toss it twice more: both heads. Now you're getting impressed. That was 4 heads in a row. What is the p-value now?

Each toss is 50:50 so the p-value changes from 0.5 on the first to 0.25 on the second toss, then 0.125 on the third, then halve that again to 0.0625 after the fourth toss. Could that be chance? It could, but the likelihood is falling. When do we reach the point where we call it 'significant'? Arbitrarily, **significance is reached once the p-value is**

<0.05. In plain English: *There is less than a one in twenty (or 5%) likelihood that this result happened by chance.*

In our example, we need a 5^{th} toss to be heads to get a p-value of half of 0.0625, which is 0.03125 (let's call it 0.03).

Does it prove that the coin ALWAYS lands on heads? No, it doesn't. It means there's a one in 32 chance I got lucky!

Lower p-values are sometimes used, such as p <0.01 (or even p<0.001), indicating greater statistical significance.

Statistical significance does not necessarily mean that a result is clinically significant.

1.3. Confidence intervals

It's often difficult to determine an exact number for a result. The confidence interval represents the range of values between which the true result lies. Most medical academic papers use a 95% confidence interval, which means that there's a 95% chance that the result is somewhere between the upper and lower limit.

A new drug, 'Cardiocalm' was found to lower systolic blood pressure by a mean -10mmHg in a study of 10 000 patients (95% CI -12mmHg to -8mmHg). Here, we can see that in 19 out of 20 cases, the blood pressure drops by somewhere between 8 and 12mmHg. There will be one in twenty where it is greater or less than this amount.

A competitor rushes out a similar drug called 'Tranquipulse', which it claims is superior. It lowered systolic blood pressure by a mean of 15mmHg in a trial involving 10 patients (95% CI -40mmHg to 10mmHg).

The confidence interval in this example is wide as there were only 10 participants. A more powerful study (with a greater number of participants) will usually result in a

narrower confidence interval. We also note that this confidence interval goes from minus 40 to plus 10. It crosses zero, the point of no effect. This means that the real outcome of this trial could be that the medication actually *increases* blood pressure.

Results which have a confidence interval crossing zero are not significant.

1.4. Relative and absolute risk

The relative risk is how many times more likely it is that an event will occur in the treatment group than the control group. A relative risk of 1 means 'no difference'. The relative risk is calculated as 'risk in treatment group/risk in control group'.

The absolute risk is the actual event rate in the group. The absolute risk reduction is the risk of the event in a control group – the risk of the event in the study group. If the absolute risk reduction is zero then there is no difference between the two groups.

***Pitfall**: relative risks can often appear striking, but must be interpreted within the context of the baseline event rate.*

The DailyWail screamed: "Limb loss horror! Cardiocalm DOUBLES risk of amputation." While this may sound alarming, the absolute background rate was merely 1:100 000, rising to 2:100 000 (= 1:50 000), which still represents a notably low risk. Beware relative risks. Instead, seek the absolute values.

1.5. Association or causation?

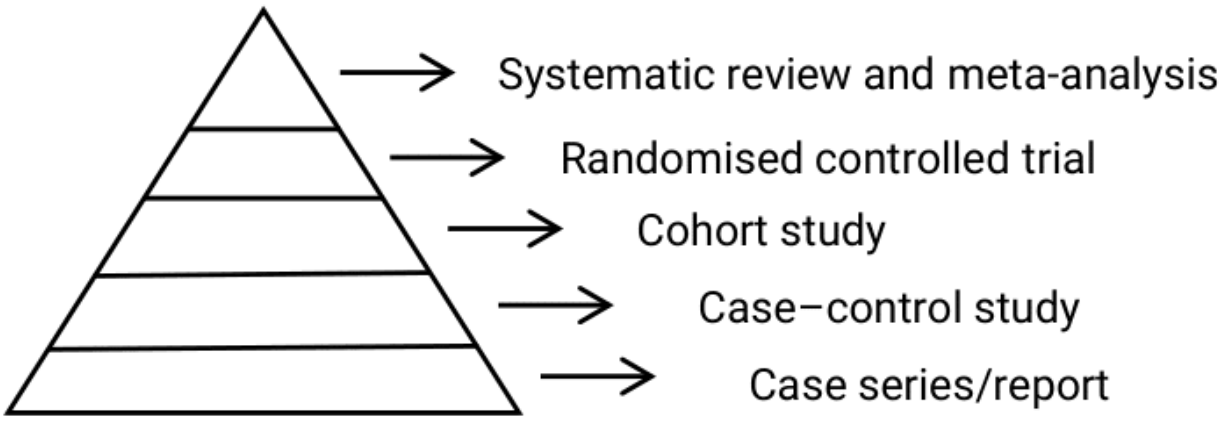
Some things cause another, e.g. snakes and snake bites. Others just happen to be observed together, e.g. umbrellas and rain, fireworks and darkness, paracetamol and chest pain.

Beware implying causation which has not been proven. Does paracetamol cause chest pain, or do people with chest pain take paracetamol? Observational trials show associations. Causation can only be established with a control group, usually in a well-conducted double-blind, randomised controlled trial.

Nevertheless, there are situations, particularly when practical or ethical constraints are in place, where causation may be inferred based on a substantial body of observational evidence. A noteworthy example is smoking causing lung cancer. For more about this, we recommend 'Bad Science' by Ben Goldacre.

1.6. Hierarchy of evidence

The hierarchy of evidence can help us decide how useful or important a study is. A number of different study designs are used in medical research, and the choice of study design relates to the question we are trying to answer. The hierarchy of evidence is often expressed as a pyramid, and the systematic review and meta-analysis sits at the very top (Hosp Pediatr (2022) 12 (8): 745).



Systematic reviews and (network) meta-analysis

A **systematic review** draws together as much evidence as possible on a topic (published and unpublished), selects the best evidence and attempts to draw conclusions.

A **meta-analysis** uses quantitative methods to pool results from similar trials. This leverages a larger sample size, which can potentially yield a statistically significant result. Beware: a good meta-analysis relies on high-quality, congruent data, i.e. source trials all investigated the same aspect, meaning you are 'comparing apples with apples'!

A **network meta-analysis** compares three or more interventions using a network of studies, e.g. which combinations of analgesics work best in low back pain: paracetamol + ibuprofen, ibuprofen + codeine or codeine + paracetamol?

Randomised controlled trial

A randomised controlled trial can be used to establish causality, and is the best individual study design for assessing drug treatments or interventions. One treatment is compared with placebo or another treatment. Studies of drugs are usually 'double-blinded', meaning that neither the patient nor the researcher know who is receiving the study drug. Randomisation and blinding both reduce bias in the study.

Cohort study

Cohort studies often involve large numbers of patients and are observational in design, meaning there is no intervention. They may be prospective or retrospective. In a prospective study, patients are recruited and then followed over many years; this is useful to study the effect of exposures, for example the relationship between smoking and lung cancer. Retrospective cohort studies look back at past exposures. A well-conducted cohort study can collect evidence towards causality.

Case–control study

A case–control study is useful for studying rare diseases or outcomes because cases can be selected and then matched with controls to study the effects of past exposures.

Case series/report

A case series or case report is considered the lowest form of evidence, but can be particularly useful for rare outcomes or events and for noticing new patterns of disease. The COVID pandemic was first recognised through a case series of severe pneumonia in Wuhan, China.

1.7. Sensitivity and specificity

Sensitivity is the 'pick-up rate'. What proportion of the people with a disease are detected by the test?

Specificity is how good the test is at ONLY being positive in those with the disease. This is often explained in terms of how good the test is at detecting the people who don't have the disease and test negative.

1.8. True and false positive/negative

True positives have the disease and test positive.

False positives do NOT have the disease and test positive.

True negatives do not have the disease and test negative.

False negatives HAVE the disease and test negative.

1.9. Positive and negative predictive value

The positive predictive value is the proportion of people testing positive who have the disease.

The negative predictive value is the proportion of people testing negative who do not have the disease.

1.10. Number needed to treat (NNT) and number needed to harm (NNH)

The NNT(H) tells us how many people we need to treat for one to benefit (harm). For NNT, a smaller value infers greater chance of success: an NNT of 1 means that everyone treated receives the benefit. An NNT of 10 means we treat 10 and one benefits. It is calculated as the inverse of the absolute risk reduction (1/ARR).

1.11. Endpoints

A **hard endpoint** is an objective, measurable outcome. Death is often quoted as the ultimate hard endpoint.

A **surrogate marker** is an endpoint which is assumed to be related to a desirable hard endpoint, e.g. drug A improves coronary blood flow. The authors may subsequently talk about reduced rates of MI or CV death, but this has not been proven.

A **composite endpoint** is a collection of endpoints which, taken together, are presented as being significant, e.g. a composite endpoint may include MI, LDL-cholesterol and cardiovascular death. The issue is that the endpoints are not equal. We can't compare

death with LDL level. Composite endpoints allow researchers to bury bad news. The drug may not reduce death, but, if we included LDL level, we can package 'reduced death' into the proven endpoint.

1.12. Percentages and percentage points

Is a percentage the same as a percentage point? We must be careful and clear. An example to illustrate:

- Drug A reduces the chance of a disease from 20 in 100 (20%) to 10 in 100 (10%).

By how much does drug A reduce the chance of the disease?

The answer is in the box at the end of the article!

1.13. Type 1 and type 2 error

These terms help you sound very clever, but they are deliciously simple:

- A type 1 error is a false positive, e.g. FIT positive with no pathology.
- A type 2 error is a false negative, e.g. FIT negative with colorectal cancer.

1.14. Absence of evidence is not evidence of absence

Do not confound these two concepts. Consider these two examples, only one of which is true:

- '**There is insufficient evidence to support the effectiveness of glucosamine for hip**

osteoarthritis.' This type of statement suggests an absence of high-quality evidence, and is usually accompanied by a call for more research.

- **'Evidence proves homeopathy ineffective.'** This is a positive statement drawn from high-quality evidence. Further research will not be helpful because we have evidence of absence of effect.

This concept is clinically relevant. NICE makes positive evidence-based statements, e.g. offer metformin to adults with type 2 diabetes. Conversely, NICE makes negative recommendations where appropriate, e.g. do not start opiates for chronic primary pain. NICE sometimes chooses not to comment if the evidence about an intervention is insufficient.

1.15. Placebo, nocebo and drug trials

You are likely familiar with the **placebo effect**, where a sham intervention has a measurable psychological or psychophysiological effect (Am J Psychotherapy 1964;18:73). Placebo is the benchmark (the control) against which new interventions are tested. Placebo is a fascinating phenomenon:

- Four sugar pills work better than two sugar pills in healing duodenal ulcers (Br J Clini Pharmacol 1999;48:853).
- Branded placebo tablets work better than unbranded placebo tablets (Health Psychology 2016;35:187).
- In hypertrophic cardiomyopathy, a pacemaker reduces chest pain, palpitations, dyspnoea and outflow obstruction gradient, even in those whose pacemakers are not yet switched on (Am J Cardiol 1999;83:903).
- Red analgesics are superior (this often-cited trial only had 22 patients, with only 5 of these taking the red pill – maybe an example of a small study gaining repeated citations, momentum and becoming embedded) (BMJ 1974;4:196).

Less well-known is the **nocebo effect**, the negative counterpart of placebo. Here, negative expectations induce perceived or measurable adverse effects (Psychosom Med 2011;73:598):

- Nebulised water causes airflow limitation in asthma if participants undergo 'bronchoconstrictive suggestion' (a prior warning that the aerosol may induce chest tightness) (Br J Clin Psychol 1986;25:173).
- Red light visual stimulus intensifies pain compared with green light (Pain 2013;154:1312).
- Statins cause more muscle aches when participants know they are receiving a statin (Lancet 2017;389:2445).

Drug manufacturers may choose to test their new drug against placebo. Watch out for this. A comparison trial using the old drug as an active control is the better test. This brings us nicely to non-inferiority trials.

1.16. Non-inferiority trials

Many trials are superiority trials, i.e. 'is this new drug better than the gold-standard drug (or placebo)?'. Sometimes, it is not ethical to test a new treatment against placebo, e.g. DOACs in AF to prevent stroke. In this situation, the early trials seek to prove that DOACs are not inferior to warfarin. Non-inferiority trials do not demonstrate superiority.

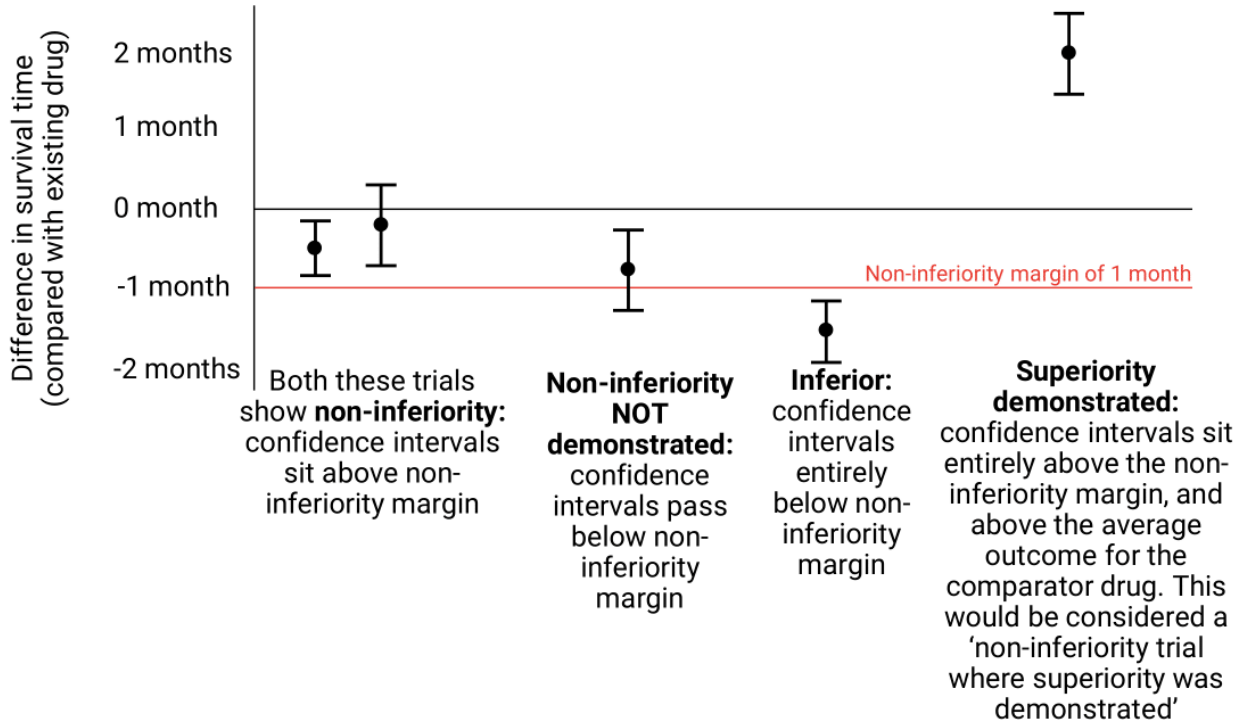
There are potential problems with non-inferiority trial design:

- We are assuming that the current best treatment is effective.
- A 'margin' for equivalence has to be established because the two treatments will not have identical efficacy. This margin could be set nice and wide, making almost anything look 'as effective'. It is therefore crucial that an appropriate non-inferiority margin is set because this will dictate the conclusion of the trial. The

choice of non-inferiority margin must ensure that the clinical difference between the two treatments is minimal (Contemp Clin Trials Commun. 2019;100454).

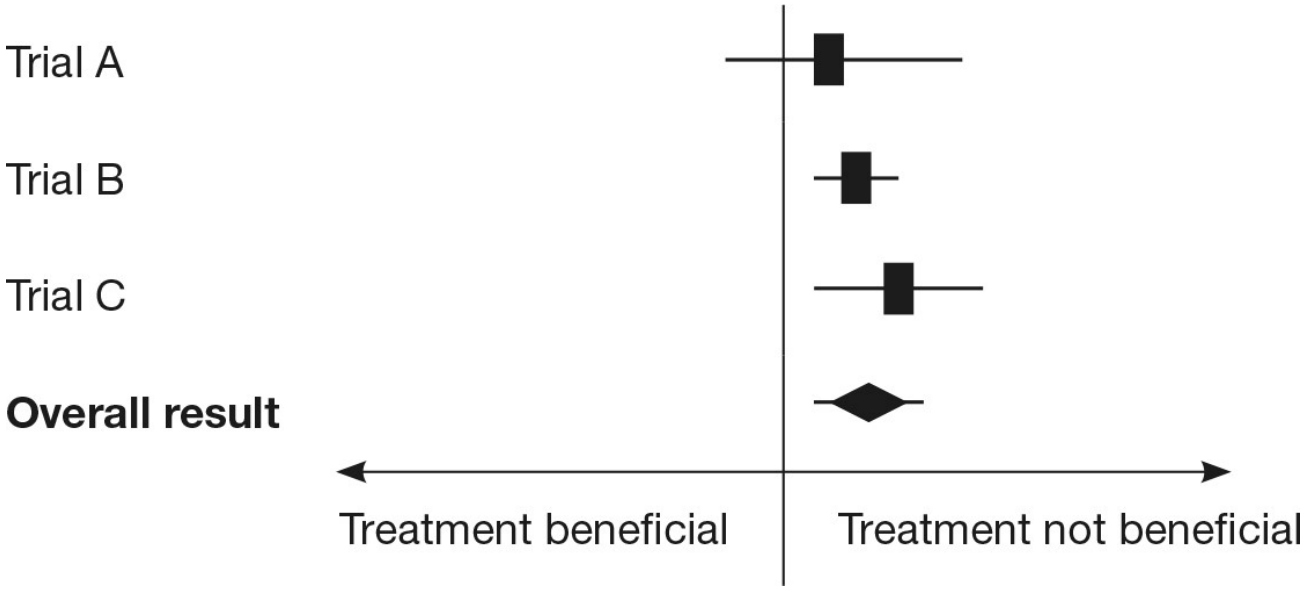
Always remember to look at the confidence intervals because they determine if a treatment is non-inferior or not. Why? Let's look at a trial of a new drug to see how it affects life expectancy.

- Let's say our non-inferiority margin was set at 1 month. This means that to be considered non-inferior to the current drug, the new drug must have an average life-expectancy no worse than 1 month less than the drug we are comparing it with.
- There are FOUR possible outcomes:



1.17. Forest plots

Above, we talked about meta-analysis as a way to leverage a larger sample size by pooling results. The results from a meta-analysis are often presented in a forest plot, which looks something like this:



Each square represents the outcome of a trial, with the horizontal line over which it sits representing the confidence intervals. The vertical line indicates where a treatment changes from being beneficial (to the left of the vertical line in this example) to not beneficial (to the right of the vertical line in this example). At the bottom, all the trials are combined and this is shown by the diamond.

In our example, the diamond is to the right of the vertical line, showing no benefit. You know this is statistically significant because the confidence intervals (the horizontal line over which the diamond sits) do not cross the vertical line.

Watch out! The 'treatment beneficial' side is not always on the left so read the labels on each forest plot carefully.

1.18. Prior probability, base-rate fallacy and the false-positive paradox

Buckle up, it's time for the grand finale, and it's a concept that will help your clinical practice. Prior probability is also called pretest probability. To illustrate, let's consider a fictional scenario.

The PSA2 test can detect prostate cancer. It has a sensitivity of 100% and a specificity of 95%. That sounds amazing. We start using it to test patients aged 40–50y with lower urinary tract symptoms (LUTS). The prevalence of prostate cancer in this age group is 1 in 1000 (we made this figure up, but go with it). If we find a positive PSA2, what are the chances of prostate cancer?

We do 1000 tests. The test picks up the one prostate cancer (true positive) and, for each remaining set of 100 tests done, we get 5 false positives (reflecting the 95% specificity). Therefore, after 1000 tests, we have just under 51 positives (1 true positive and marginally fewer than 50 false positives from the remaining 999). The answer is therefore 1 in 51. That's not a great diagnostic test – the 'accuracy' or positive predictive value is around 2%, despite first impressions being 'amazing'.

In contrast, we do the same thing with 1000 symptomatic male patients aged >75y with newly progressive, obstructive LUTS. Here, our pretest probability is much higher, say 1 in 10. How does our test perform now?

We do 1000 tests. The test picks up the 100 prostate cancers (the true positives) and, for each remaining set of 100 tests done, we get 5 false positives. We obtain 145 positives (100 true positives and 45 false positives). Here, the test performs much better. The accuracy of the very same test is now around 70% due to the higher 'prior probability'. A positive PSA2 result is much more likely to be due to prostate cancer in the older age group than in the younger group. Most positive results in the younger group will be false positives.

Base-rate fallacy occurs when we ignore the prevalence (the base rate) and focus only on the case presentation, e.g. choosing to do a PSA on a 20-year-old with LUTS. Understanding this helps us avoid doing unnecessary tests!

The false-positive paradox describes the situation when false positives are more numerous than true positives due to the low base rate. If we repeated the PSA on a series of 20-year-olds, how many false positives would we encounter before finding a true positive?

In conclusion, we must think not only about the test itself, but also about the characteristics of the population to which we apply it.

Statistics for the GP trainee

- Mean (average) is commonly used, but median is not skewed by an outlying value.
- The p-value is the decimal representation of the percentage chance of something happening.
- Confidence interval is the range in which the true value is thought to lie with a specified certainty.
- Absolute values are exact, independent of other data, while relatives compare datasets.
- Observational studies do not prove causation, only association.
- Meta-analyses and systematic reviews are high-level evidence BUT rely on high-quality inputs (garbage in = garbage out). NNT and NNH are practical statistics, and are useful in explaining risks and benefits to patients.
- Beware composite endpoints and surrogate markers.
- Absence of evidence is not evidence of absence.
- Does an intervention reduce your chance of an outcome by 5% or 5 percentage points?
- Is the trial comparing the new treatment against placebo or against current best treatment?
- Non-inferiority trials are not designed to prove superiority.
- Be mindful of the population to which you apply a test.

Practise your explanation of a common scenario, e.g. anticoagulation to prevent stroke in AF or statins for prevention of CVD. Consider videoing yourself. Is this easy to understand for your patient? Can you use a resource as a decision aid? Percentages and percentage points answer: the drug cuts the risk of the disease by 50%, or by 10 percentage points. This is a practical example of absolute vs. relative risk.

Useful resources:

Websites (all resources are hyperlinked for ease of use in Red Whale Knowledge)

- CEBM Oxford
- GP Evidence (this is a fantastic site with resources to help you explain common concepts to patients, e.g. anticoagulation in AF, statins in primary prevention)
- The NNT (contains reviews, set out in topics, with traffic-light ratings for effectiveness)

Books

- *How to Read a Paper*, Greenhalgh, T (6th Ed, Wiley-Blackwell)
- *Medical Statistics Made Easy*, Harris, Taylor (3rd Ed, Scion Publishing)

Videos

- YouTube – Battling bad science, by Ben Goldacre

any such liability cannot be excluded at law, we do not accept any liability for loss of any type caused by reliance on the information in these pages.

Here is the link to our terms of use.