

# P Values, Statistical Significance & Clinical Significance

When looking at the results of a research study, a practitioner has to answer two big questions:

1. **Were the results due to chance?**
2. **Are the results big enough to matter to a patient?**

## P values and Statistical Significance

When looking at the results of a study, a natural question is—*is it likely that the reported results were due to random chance alone?*

A quick and simple item to look at is the *p value*. The p value tells you how *probable* the results were due to luck.

- .10 means that there is a 10% probability the results were due to random chance.
- .05 means that there is a 5% probability that the results were due to random chance.
- .001 means that the chances are only 1 in a thousand.



In health care research, it is generally agreed that we want there to be only a 5% or less probability that the treatment results, risk factor, or diagnostic results could be due to chance alone.

When the p value is .05 or less, we say that the results are *statistically significant*. Results that do not meet this threshold are generally interpreted as negative.

## Clinical Significance/Importance

The results of a study can be statistically significant *but still be too small to be of any practical value*. This is of great importance to physicians when looking at research evidence.

Various quantitative measures are used to decide whether a treatment effect is *large enough to make a difference to a patient or doctor*. How much decrease in pain is large enough to matter? How much improvement in function is enough to make a treatment worthwhile? How many additional minutes/months/years of extended life make a cancer treatment worthwhile?

To a large degree, this is a subjective judgment made by the physician (or the patient). Usually the extremes are easy to recognize and agree upon. If a treatment on average will only decrease a patient's pain intensity ½ point on an 11 point scale, most of us would agree that we should try to find a better treatment option. If on the other hand, patients get 90 or 100% pain relief, we can all agree that this is an effective, worthwhile treatment (setting aside cost and side effect considerations).

But what would constitute the smallest amount of improvement that would still be considered worthwhile? After all, we want our treatments to make a difference. This is tricky. The term for this is *minimal clinically important difference* (MCID). To a large degree, practitioners must use their own judgment in deciding how much is enough. Besides using their own judgment, they sometimes can get guidance from various sources.

Sometimes the researchers doing the study will explicitly state what this minimal amount of clinically important improvement is; sometimes previous research has been done to determine this threshold. The terms to look for in a study are whether the results were *clinically significant*, *clinically important*, or met the required *MCID*. Many times it is dependent on the method or tool used to measure improvement.

Look in both the RESULTS section and DISCUSSION section of a study. See what outcome measures were used and how much they improved. If it is therapy study comparing two types of treatment, don't just look at the comparisons between the two treatments, *look to see how much patients improved compared to their baseline*. After all, one treatment might be statistically more effective than the other, but neither might end up improving the patient much. Unfortunately, sometimes this information is hard to find and is not highlighted in the ABSTRACT or the DISCUSSION or the CONCLUSION. Sometimes it is buried in the RESULTS section, sometimes found only in tables or graphs.

When you can find the absolute amount of improvement in each outcome measure, you can then decide for yourself if the improvement looked very large and you can sometimes cross reference it with other sources to decide whether it met the MCID.

You often can find a suggested MCID in the UWS CSPE protocols (on pain severity and the various questionnaires).

Here are some examples:

Condition	Outcome measure	Suggested MCID
Low back pain	0-10 pain scale	1-2 points or 30% reduction
Musculoskeletal injury	PSFS	2 for average of 3, 3 points for one item
Low back pain	Oswestry questionnaire	4-6
Low back pain	Roland Morris questionnaire	2-5
AROM	Observation, goniometer	Around 20% improvement (although it would further be influenced by the specific joint in the body and the amount of improvement that might impact a patient's individual job demands)

When systematic reviews report on multiple studies, they may combine the results and report them in terms of overall *effect size*.<sup>1</sup> Since effect size numbers do not make intuitive sense, you can consult a general guideline as follows:

- 0.2 = a small treatment benefit
- 0.5 = moderate size treatment benefit
- 0.8 = a large treatment benefit

More and more therapy studies are reporting clinical improvement by citing the number of patients that would need to be treated to have one successful outcome that would not have occurred without that therapy. This is called the number needed to treat (NNT). An NNT of 1 would be the perfect treatment. Everyone treated got important improvement and would not have improved without the treatment. Generally speaking NNT for therapies should be in the single digits (e.g., 1-10). Even then, one must use one's judgment as to whether the NNT is low enough considering cost, side effects, and the harm that might result from not being successfully treated. For preventive measures, NNTs are often in the double digits. For more information, go to the EBP boot camp document Number Needed to Treat.

---

<sup>1</sup> The effect size is a somewhat complicated creature. It is not a likelihood ratio or odds ratio. It is a method of demonstrating how much better one intervention group did compared to another. It is calculated by taking the difference between group means divided by the standard deviation. The larger the number, the stronger the beneficial effect.

## Bottom line

Don't just look at the p value. Try to decide if the results are robust enough to also be clinically significant. This is important enough that it should always be considered by the practitioner (and reported by the student when constructing a CAT).

We might have a wonderful new treatment that can reduce someone's pain 5% on average with a p value of .0001. This means we are really, really sure that the results are not accidental-- the improvement is really due to the therapy and not just chance. On the other hand, who cares about a treatment with such a paltry effect? The results are *statistically significant* but not *clinically significant*.

---

R LeFebvre, DC  
reviewed by Mitch Haas, DC  
2/15/11