

A GUIDE TO STATISTICAL SIGNIFICANCE IN EMS RESEARCH

Understand a few key concepts to better comprehend research and its clinical significance

By Lawrence H. Brown, PhD

When conducting research, it is impractical (and usually impossible) to study every person with a disease or problem—what researchers call the *target population*. Imagine the resources required to study every heart attack patient in a year or every car crash victim over a decade! Instead, researchers select a sample of patients to represent the larger target population.

Although researchers care about the people in their studies, what they really want to know is what those people (the *sample*) can teach them about the larger (*target*) population. One statistical technique researchers use to describe what they learn about the target population

from a sample is *confidence intervals*. A confidence interval takes data for some measure obtained from a sample and then calculates what that measure probably looks like in the target population. Typically researchers use and report confidence intervals of 95% (95% CI).

For example, if researchers are studying the heart rates of trauma patients, they might find an average heart rate in their sample of 102 bpm. Using that average, the standard deviation, and the number of people in their sample (find the formula at www.wikihow.com/Calculate-Confidence-Interval), they might calculate a 95% CI of 98–106. That means in this sample of trauma patients, the average heart rate was 102, and in the

target population of all trauma patients, the researchers are 95% sure the average heart rate is somewhere between 98 and 106.

Ideally the 95% CI is narrow enough that there is no practical difference between the measure found in the sample and the probable range of that measure in the target population. If the confidence interval is wide—for example, a sample average of 102 but a 95% CI ranging from 52 to 152—then the sample doesn't provide a very clear indication of what's going on in the target population. The width of a 95% CI is driven by the number of subjects in the sample and the natural variation in whatever's being measured. Confidence intervals can be calculated for almost every type of measure (averages, medians, percentages, ratios, etc.).

Comparing Subgroups

Confidence intervals can also be used to compare two or more subgroups within a sample. For example, imagine researchers studying a target population of congestive heart failure patients select a representative sample of patients. They administer nitroglycerin to half the sample (the *intervention group*) and a placebo to the other half (the *control group*).

If 50% of the patients in the intervention group and 60% of the patients in the control group require ICU admission, then there is a difference in admission rates of

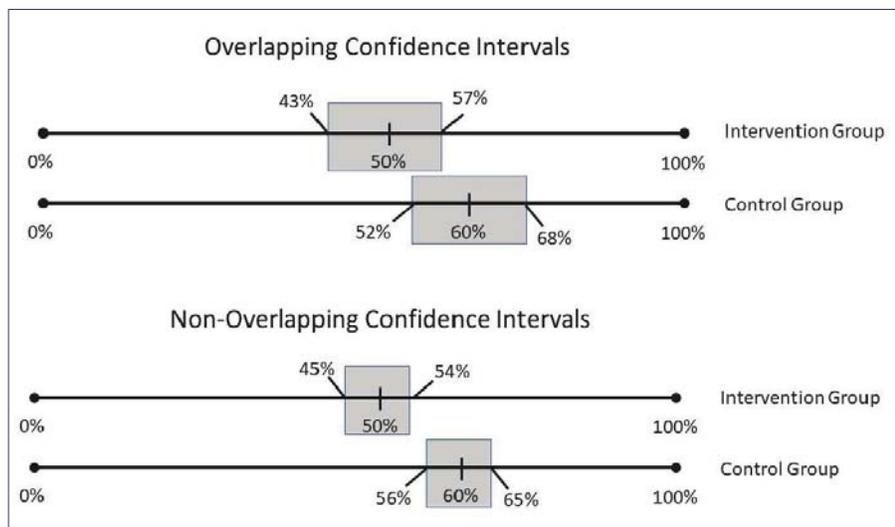


Figure 1: Overlapping vs. non-overlapping confidence intervals

those two subgroups within the sample. By calculating a confidence interval around the observed proportions, researchers can estimate the effect in the target population.

If the 95% CI for the intervention group is 43%–57%, then investigators are 95% sure that between 43% and 57% of congestive heart failure patients in the target population who receive nitroglycerin will require ICU admission. If the 95% CI in the control group is 52%–68%, those two confidence intervals overlap (see *Figure 1*).

That is, ICU admission rates in the target population could be the same in both subgroups. On the other hand, if the 95% CIs for the two subgroups do not overlap, we would be 95% certain the ICU admission rates for two subgroups differ in the target population. As with descriptive uses, this comparative use of confidence intervals works for almost every type of measure.

Using Statistical Tests

Another way researchers compare subgroups within a sample is through statistical testing. There are many different tests, and the chosen test depends on the type and characteristics of the data being analyzed. Whatever test is used, the most commonly reported measure produced by statistical tests is the *p-value*.

The *p-value* represents the probability of finding some statistical test result simply by random chance. More practically, the *p-value* is the probability of finding an observed difference (or association, or whatever is being measured) in two sub-

groups of a sample if those subgroups truly represent the same target population.

Using the congestive heart failure example above, the question is whether patients who receive nitroglycerin and patients who do not receive nitroglycerin are all just part of the same big target population (i.e., the differences are just random variation), or are they really two separate target populations (i.e., there are true, consistent differences; see *Figure 2*)?

Researchers usually use a threshold (called an *alpha value*) of 5% to establish statistical significance. If the statistical test generates a *p-value* less than the threshold *alpha value*, that means there is less than a 5% chance that the data come from two subgroups in a sample representing one big target population with some natural variation. Instead the data probably represent two samples from two truly different target populations.

Importantly, *p-values*—like confidence intervals—are strongly influenced by the number of subjects included in an analysis. If a study reports a *p-value* greater than or equal to 0.05, we are left to wonder whether the data for the two subgroups truly represent the same larger target population or whether the study was just too small to detect that the two samples actually represent two different target populations. The power of a study is the probability that it will detect a difference if one exists. Researchers typically design studies to have at least 80% power, but this is not always possible.

Studies with very large numbers of subjects (especially retrospective analyses of databases with thousands of subjects) have extreme power and can produce *p-values* less than 0.05 even when differences between two subgroups are quite small. For these kinds of analyses, researchers sometimes use a more conservative *alpha value* threshold of 0.01 to establish statistical significance.

Clinical Significance

Statistical tests and *p-values* are measures of probability, not the size or strength of a difference or association. Whether a difference between two subgroups in a

sample is practically meaningful is a question of clinical significance. If a finding isn't clinically significant, it doesn't really matter whether it's statistically significant.

Clinical significance requires professional judgement informed by experience and practicalities. A study of an intervention that reduces mortality from 18% to 15% with a *p-value* of 0.003 likely has less practical impact than one of an intervention that reduces mortality from 18% to 5% with a *p-value* of 0.038—even though the first study produces a much smaller *p-value*. Similarly, a study of an intervention that reduces admission rates from 23.7% to 23.4% would have little practical significance even if the *p-value* were 0.001. Thus researchers think of statistical significance as a binary yes ($p < 0.05$) or no ($p > 0.05$) concept and avoid describing findings as “slightly significant” (e.g., $p = 0.048$) or “very significant” (e.g., $p = 0.001$).

Putting It All Together

By combining a basic understanding of sampling, confidence intervals, *p-values*, and statistical and clinical significance, readers can better judge studies they read and understand their clinical importance. They can extrapolate data in a study to the target population of their own patients; they can determine the probability that a study's effect is simply a function of random variation within the target population; and they can determine whether the findings of the study are clinically or practically meaningful. This is the crux of analyzing research findings. 🌐

ABOUT THE AUTHOR

 **Lawrence H. Brown, PhD**, a former paramedic, is an associate professor and director of research education for the Emergency Medicine program at the University of Texas' Dell Medical School.

 This article was produced in partnership with the Prehospital Care Research Forum at UCLA. Visit: <https://www.cpc.mednet.ucla.edu/prcf>


MORE ONLINE!
 For a guide to searching the scientific literature visit www.emsworld.com/article/220373.

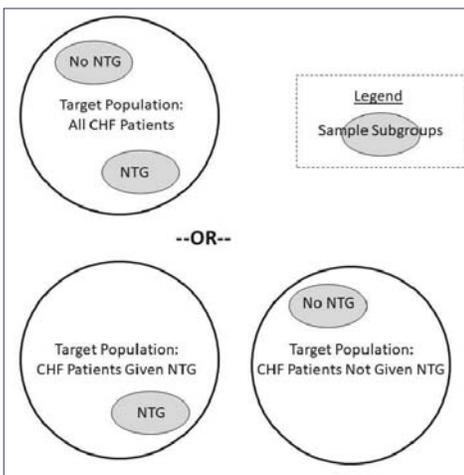


Figure 2: One vs. two target populations