Understanding equivalence trials (and why we should care)

Jacques S. Lee, MD

Introduction

In the April 2000 issue of *CJEM*, Wood and colleagues compared titrated IV meperidine to single-dose IV ketorolac in patients with acute renal colic.[1] They achieved successful pain relief in 72% of patients who were administered ketorolac vs. 64% of patients who were administered meperidine, yet concluded "equivalence." Why? To compound the confusion, they indicate that they tested the null-hypothesis that meperidine is 20% better than ketorolac. But isn't that backwards? Isn't the "null hypothesis" that there is no difference between groups?

What we have here is a different animal — an equivalence trial — and you'll be seeing many more of these in the future.

What is an equivalence trial?

Medical science is preoccupied with improving existing treatments, so most clinical trials are designed to show that a new treatment is better than placebo or better than standard therapy. And rightly so, otherwise the textbooks would still cite "leeches" and "eye of newt" as time-honoured therapies. But when proven therapies already exist and a new treatment is unlikely to have superior efficacy (but may be safer, easier to use or less expensive), we sometimes want to know whether the new treatment is "as good as" standard therapy. Equivalence trials are used to prove that two treatments work equally well, in contrast to traditional "superiority trials," which set out to show that one treatment is better than another.

Defining "equivalence"

Since no two treatments are exactly equal, we must define what difference in effects would be clinically important. This clinically important effect size (delta) will differ depending on the condition being studied. In thrombolytic trials, an absolute mortality difference of 1% is often defined as the minimal clinically important difference. In most other situations (e.g., migraine headache), a 20% difference would be considered important and a 1% difference meaningless.

When looking for small differences in treatment effect, large patient samples are required. This is because it is difficult to distinguish small treatment effects from background sampling error. On the other hand, large treatment effects may be statistically significant despite small sample sizes because sampling error rarely causes large differences in study outcomes. Norman and Streiner[2] have referred to this as the"signal-to-noise ratio." To illustrate, it took over 40,000 patients to show that tPA was statistically superior to streptokinase for acute myocardial infarction (AMI) (6.3% vs. 7.2% mortality; $p = 0.001$), but only 98 patients to show that dexamethasone was superior to placebo for migraine headache (18% vs. 45% headache recurrence; $p = 0.005$).[3,4] This is because the effect size was much larger in the latter study.

Hypothesis testing

After defining what we think would be a clinically important outcome difference, how can we prove that Treatment

Division of Emergency Medicine, University of Ottawa, Ottawa, Ont.

Received: Feb. 13, 2000; final submission received: Apr. 21, 2000; accepted: May 13, 2000

A is equivalent to Treatment B? Nearly all of us have been exposed to the concept of classical hypothesis testing* (although most deeply resent the experience). Tapping into these repressed memories, we remember that the first step is to formulate a null hypothesis: "Treatment A has the same response rate as Treatment B." We then perform the study and determine whether the two treatments did, indeed, produce similar response rates. If our study shows a statistically significant difference, we reject the null hypothesis and accept the alternate hypothesis (that one treatment is better than the other).

Equivalence trials reverse this logic. In an equivalence trial, the "null hypothesis" states that Treatment A (standard therapy) is *better than* Treatment B (usually the newer, easier, cheaper or safer agent) by the predefined clinically important difference. If our data subsequently reject this null hypothesis (of "non-equivalence") then we can accept the alternate explanation that the two treatments work equally well.

So what sadist came up with this twisted scheme? Were lawyers and politicians behind it? Why not keep it simple and use the traditional study design? Why not? Because, to paraphrase Churchill, "For every complex problem, there is a simple solution. And it's the wrong solution." The fact is, if we attempt to prove equivalence using a superiority design, we are destined for trouble, as the following example illustrates.

Echinacea for asthma

Consider a study in which 30 asthmatic patients were randomized to receive either echinacea or salbutamol. After 5 minutes of therapy, 4 of 15 echinacea recipients (27%) and 7 of 15 salbutamol recipients (47%) reported improvement, with no significant difference between groups ($p = 0.55$). The authors stated that, with a sample size of 30 patients, the study had 80% power to detect a 50% difference in response rates. They therefore concluded that echinacea is as effective as salbutamol for asthma exacerbations.

Thirty patients doesn't seem like many, yet these authors reported that their study had 80% power. Doesn't this mean they had enough patients to rule out important differences between the treatments? The answer is No. Their study had 80% power, but 80% power to detect a very large (50%) outcome difference, which means that clinically important differences (e.g., 40%) could go undetected. When evaluating power statements, readers must consider what effect size the investigators defined as clinically important.

The echinacea investigators used a classic superiority design. Its results could be accepted if they showed one agent to be statistically better than the other. But they didn't, and with a negative study, all bets are off. Failure to show superiority doesn't prove equivalence, and just because the study failed to show that echinacea is better than salbutamol, we can't conclude they are the same.

We must be suspicious of studies like this, because using a superiority study design when an equivalence design is more appropriate rewards sloppiness. Note that by measuring their outcome early (at 5 minutes) and by delivering standard therapy in a suboptimal manner (e.g., not coaching or using a spacer device), the echinacea authors ensured that both treatments would appear equal — equally bad!

Enrolling too few patients is another common pitfall. The smaller a study, the less able it is to find a statistically significant difference. When trying to prove that two treatments are essentially equal, then a larger study is more convincing.

Sampling and sample size

Sampling is a critical part of study design, and the echinacea conclusions are invalid because sample size was inadequate. Remember that patients in any study represent only a sample of the larger population of interest, and that doing a study is like sticking your hand into a box of Cracker-Jacks®. If you pull out a handful of popcorn, you might conclude that there aren't any peanuts in the box. Children familiar with the true composition of Cracker-Jacks® would call this a sampling error. We can call it plain bad luck.

Studies, like diagnostic tests, suffer from 2 types of sampling error: false-positive results (type I, or alpha error) and false-negative results (type II, or beta error). Finding no peanuts in a handful of Cracker-Jacks® when there actually are peanuts in the box is an example of a false-negative study or type II error. By convention, we accept a 20% chance (beta = 0.2) that a study's results will be falsely negative due to "bad luck" or sampling error. The "power" of a study is a more macho way of saying this: A study with a 20% chance of type II error has 80% power to find the predefined clinically important difference if it actually exists.

Power is related to sample size, and small studies are prone to sampling error. The only sure way to determine the true popcorn:peanuts ratio is to study the whole population, in this case the entire box of Cracker-Jacks®. A highly skilled researcher might be willing to tackle this, but when

*Hypothesis testing is the most common method, but other statistical approaches can be used for equivalence trials. For example, precision estimation compares the confidence interval of the two treatments' response rates. If these are seen to overlap within a specified range, then equivalence can be concluded.

you're studying people with diseases, it's usually impossible to study the whole population. Fortunately, studying a large enough sample of patients can provide accurate estimates of truth. Prior to embarking on a study, researchers perform a sample size calculation to determine the minimum number of patients necessary to reduce the chance of sampling error to an acceptable level.

Statistical significance vs. clinical importance

Consider what might happen if the GUSTO investigators turned their attention to otitis media, and found, in a study of 40,000 children, that imipenem is superior to amoxicillin (cure rate 96.1% vs. 95.2%; $p = 0.002$). This difference is statistically significant but clinically unimportant. Contrast this to the findings of the asthma study, where success rates were 27% for echinacea and 47% for salbutamol ($p = 0.55$). The latter results are clinically important but statistically insignificant. When interpreting the outcome of any study, physicians should consider both the reported effect size (best estimate of truth) and the *p* value (likelihood that this difference arose by chance). If the effect size was clinically important, but the *p* value insignificant, the sample size was probably too small (underpowered study). If the effect size was clinically unimportant but the *p* value significant, the sample size was probably too large (overpowered study).

Physicians should be aware that pharmaceutical manufacturers may be tempted to design overpowered studies, in the hopes of finding small but statistically significant differences that will increase use of their product. These differences may be clinically unimportant, and one of our great challenges as physicians is to establish consensus about what are clinically important outcome differences for various disease states.

Equivalence in renal colic

In the renal colic study,[1] Wood and colleagues defined the clinically important effect size as 20% and acknowledged a 5% chance of type I error (alpha = 0.05), and a 20% chance of type II error (beta = 0.2). Based on these parameters, they calculated that 126 patients were necessary to prove equivalence. After analyzing their data, they concluded that ketorolac and meperidine are statistically equivalent ($p = 0.02$). In this case, equivalence was the target; therefore the trial was positive. Given positive trial results, our first question should be: What is the likelihood that this was a false-positive (type I error) study? (Remember that, in an equivalence trial, type I error means to incorrectly conclude equivalence, while type II error means to incorrectly conclude non-equivalence.) The *p* value (for equivalence) of 0.02 tells us that, given the difference seen in the study, there is a 2/1000 chance that meperidine is actually better by the predefined delta of 20%. If the trial had failed to show equivalence (negative trial), the authors could state, based on an 80% power calculation (beta = 0.2), that they are 80% sure that meperidine and ketorolac are not equivalent (true negative study). They could *not* conclude that one treatment was better than the other: That's not the point of an equivalence trial.

These authors avoided many of the "equivalence pitfalls" demonstrated in the echinacea study. They compared sufficient, titrated doses of meperidine to a single and, arguably, low dose of ketorolac. Their primary outcome (successful pain relief) was measured after sufficient time had passed to allow both drugs to produce their effect. To address the concern that ketorolac might relieve mild but not severe pain, they excluded patients with mild pain. Finally, as described, they specified a null hypothesis of non-equivalence. If we agree with their clinically important effect size, then we can conclude with confidence that meperidine and ketorolac have equivalent effectiveness in relieving the pain from renal colic.

Conclusion

As more treatments become available for illnesses where satisfactory therapies already exist, equivalence trials will become increasingly common. Emergency physicians will need to understand how to interpret published results of equivalence trials, and how to recognize the specific pitfalls that can arise in their design and execution.

References

1. Wood VM, Christenson JM, Innes GD, Lesperance M, McKnight RD. The NARC (Nonsteroidal Anti-inflammatory in Renal Colic) Trial. Single-dose intravenous ketorolac versus titrated intravenous meperidine in acute renal colic: a randomized clinical trial. CJEM 2000;2:83-9.

2. Norman G, Streiner D. Biostatistics: the bare essentials. St. Louis: Mosby; 1994. p. 42.

3. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993;329:673-82.

4. Innes GD, Macphail I, Dillon EC, Metcalfe C, Gao M. Dexamethasone prevents relapse after emergency department treatment of acute migraine: a randomized clinical trial. CJEM 1999;1:26-33.

Correspondence to: jslee@ican.net