RESOLVING DESIGN PROBLEMS IN EQUIVALENCY TRIALS

Matthew P. Fox, MPH, and LeAnne M. Fox, MD, MPH

Equivalency trials provide a practical means of demonstrating the efficacy of new interventions when ethical issues prevent comparing the intervention to a placebo.[1] Generally in equivalency trials, a new intervention is compared with a currently accepted standard treatment. The goal of such a trial is not to test whether the new intervention is better than the current standard, but rather to determine whether there is any important clinical difference between the new intervention and the current standard. When no important difference exists, then a cheaper or more easily tolerated intervention may be justified.

Equivalency trials generally require larger sample sizes than superiority trials because studies with small sample sizes that find no difference between the new intervention and a current standard are often underpowered to detect a meaningful difference. Nevertheless, even with an adequate number of subjects, equivalency trials can sometimes make 2 interventions appear equivalent, even when they are not. General issues with equivalency trials have been detailed elsewhere,[1-4] most of which concern lack of incentive for rigorous designs.[5-9] This paper explores design methods for equivalency trials that can substantially increase confidence in claims of equivalence.

Dagan and McCracken describe how, with commonly used methods, antibiotic equivalency trials often show that 2 drugs are equivalent when strong differences truly exist.[10] This effect, originally described by Marchant and colleagues and termed the "Pollyanna Phenomenon,"[11] (in reference to the overly optimistic character of Porter's novel[12]), creates an environment that can exaggerate the benefits of an inferior drug when compared with a drug that is actually therapeutically superior.

DESIGN PROBLEMS IN EQUIVALENCY TRIALS

The "Pollyanna Phenomenon" highlights how choices in clinical trial design can lead investigators to create favorable conditions for demonstrating that a new drug is as efficacious as the current standard, when, in reality, it is not as efficacious. This phenomenon is based primarily on 4 principles: 1) subjects who would experience spontaneous cure in the absence of any intervention will make an intervention appear more efficacious than it truly is; 2) when clinical criteria are used in the recruitment of subjects, those not infected with the pathogen the drug is intended to cure are likely to be enrolled; 3) using clinical criteria to diagnose whether subjects have been cured at the end of a study is never perfect, leading to misclassification of the study outcome; and 4) studies with small sample sizes may show no statistical difference in 2 interventions. This may lead to the mistaken conclusion that the 2 drugs are equivalent, when in reality the sample size does not support such a claim.

The fourth principle has been adequately described in previous work, and solutions have been proposed.[10,11] In this paper, we will examine each of the 3 remaining design problems. For each, we will describe the problem, detail how it works, and propose a solution. In the Appendix (available at www.jpeds.com), we provide mathematical discussions of these phenomena to demonstrate that they apply outside of the examples we use. Because problems 2 and 3 have the same impact (ie, subjects have no symptoms but still have bacterial disease, or have clinical symptoms but do not have bacterial disease), we treat them together.

DESIGN PROBLEM 1: INFLATION OF EFFICACY

Including subjects whose illness will spontaneously cure in the absence of treatment will inflate the overall efficacy of any drug in an equivalency study. For example, consider the extreme situation in which a placebo (which has no bacteriologic efficacy) is given to a population of subjects whose conditions would all improve even without intervention. In this case, the placebo would appear to be 100% efficacious. Because even a placebo can appear to have a high efficacy when subjects whose illness will spontaneously cure are

AOM	acute otitis media	Se	sensitivity
RD	risk difference	Sp	specificity
RR	relative risk		

From the Center for International Health and Development, Boston University School of Public Health, Boston, Massachusetts; and Division of Infectious Diseases, Children's Hospital, Boston, MA.

Supported by the U.S. Agency for International Development (USAID), Country Research Activity (GHS A-00-03-00020-00). Dr LeAnne Fox was supported by the Pediatric Infectious Diseases Society Fellowship Award funded by GlaxoSmithKline.

Submitted for publication Oct 26, 2005; last revision received Dec 21, 2005; accepted Feb 22, 2006.

Reprint requests: Matthew P. Fox, MPH, Center for International Health and Development, 85 E Concord St, 5th Floor, Boston, MA 02118. E-mail: mfox@bu.edu.

J Pediatr 2006;149:12-6

0022-3476/$ - see front matter

10.1016/j.jpeds.2006.02.036

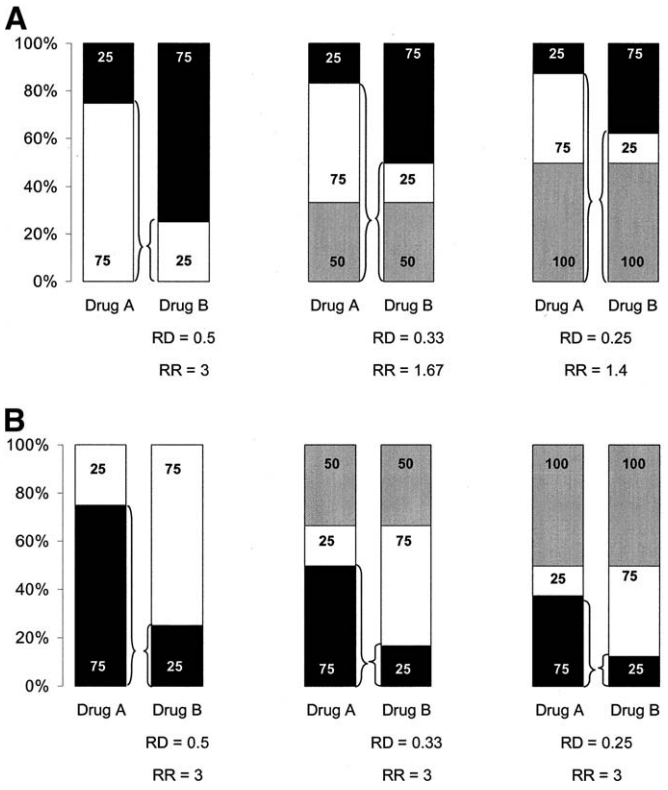

Figure 1. A and B, Demonstration of the effect of spontaneous cures on the RD and RR. Brackets represent the proportion of subjects with the outcome of interest: treatment success (**A**) and treatment failure (**B**). Black bar = failure; white bar = success; gray bar = spontaneous cures.

included in a trial, this makes comparisons difficult to interpret; 2 drugs are likely to appear to have similar efficacy despite their true efficacies being vastly different.

DESIGN PROBLEM 1: MECHANISM

Consider an antibiotic equivalency trial comparing 2 drugs to treat acute otitis media (AOM), in which we could know that none of the subjects included in our study would have cured spontaneously in the absence of treatment. In this trial, 200 subjects are recruited, 100 allocated to receive drug A and 100 to drug B. Also assume that drug A is more efficacious than drug B. In the absence of any spontaneous cures, the illness of 75% of subjects is cured when given drug A and that of 25% of subjects is cured when given drug B. This is depicted in the far left of Figure 1*A*. To compare the 2 antibiotics, we can use the risk difference (RD) or relative risk (RR). For example, our interpretation of Figure 1*A* is that the difference in risk of cure between drug A and B is 50% (RD = 0.5) and drug A is 3 times more efficacious than drug B (RR = 3).

Consider what happens to these comparisons when individuals who experience spontaneous cure are included in the study population. With randomization and a large sample size, we expect the number of spontaneous cures in each group to be equal. The middle and far right sections of Figure 1*A* show what happens to the efficacy of each drug when 50 and 100 spontaneous cures are included in antibiotic group A and B, respectively. Because the illness of these subjects will all cure, the efficacy of both drugs is increased; however, proportionally, the increase is greater for drug B than for drug A. When 50 spontaneous cures are added to each group, the efficacy of drug A increases from 75% to 83%, whereas the efficacy of drug B doubles from 25% to 50%. This effect is even greater when 100 spontaneous cures are added to each group. When comparing 2 drugs, both the RR and the RD will be biased toward equivalence (RD = 0, RR = 1) because both drugs will demonstrate reasonably high efficacy. (Formulas are available at www.jpeds.com in Appendix A.)

In reality, an investigator cannot "add" subjects whose illness spontaneously cures; however, investigators can manipulate the proportion of spontaneous cures in any study through several methods, including patient selection and by changing the timing of assessing the study outcome. For example, in delaying the study outcome (ie, choosing day 6 as opposed to day 3 for clinical evaluation of AOM cure), an investigator can increase the number of subjects whose illness would cure even in the absence of treatment.

DESIGN PROBLEM 1: SOLUTION

Now consider the situation in which the outcome of interest is changed to treatment failure of the study drug rather than treatment success. In this case, a high failure risk denotes low efficacy. Again, 200 subjects are recruited: 100 randomized to drug A, and 100 randomized to drug B. In this example, drug A is less efficacious than drug B; 75% of those given drug A will fail treatment, whereas only 25% of those given drug B will fail, as denoted in the far left of Figure 1*B*. Here, the risk of treatment failure with drug A is 3 times higher than with drug B (RR = 3), and the difference in treatment failure risk between drug A and drug B is 50% (RD = 0.5).

Again, consider what happens to the risk difference and the relative risk when individuals who experience spontaneous cure are included in the study population. When 50 and 100 subjects who spontaneously cure are included in each treatment group, the risk of failure for both drugs is reduced, as depicted in the middle and right sections of Figure 1*B*. Adding 100 spontaneous cures reduces the risk of failure for drug A from 75% to 37.5% and for drug B from 25% to 12.5%. Thus, whereas the true difference between drug A and B is 50%, it will appear to be only 25% (RD reduces from 0.5 to 0.25), biasing the comparison toward equivalence. However, as the risk of failure for each drug is reduced proportionally (in this case by half of its original value), drug A will still appear to have 3 times the risk of failure of drug B (RR = 3). Therefore, in interpreting the RR we would correctly conclude that drug A has 3 times the risk of failure as does drug B.

In contrast to using treatment success as the outcome, using treatment failure as the outcome biases the RD toward equivalence, but has no effect on the RR. Therefore, designing a trial in which the outcome of interest is treatment failure