When Is a Confirmatory Randomized Clinical Trial Needed?

Donald A. Berry*

Few clinical trials show important advantages for one therapy over another. When one does, a confirmatory trial will help decide whether the observation in the first trial was real. The need to replicate experimental results is fundamental in science. Systematic differences occur across clinical trials for a variety of reasons: differences in patient populations (geographically and over time), different patient referral patterns, differential application of entry criteria by investigators, and so on. Results vary across trials, just as they vary across patients within a trial, although intertrial variability is usually less than interpatient variability. If the differences among trials affect all treatment arms similarly, then treatment differences may be similar in different trials, but trial settings may well interact with treatment.

The need for replication suggests that the answer to the question in the title is "Always." But carrying out a confirmatory trial has implications other than scientific ones. First, trials require commitments of time and resources. More important, the original trial suggested that the control therapy is inferior, and randomly assigning patients to receive a possibly inferior therapy is ethically questionable. Randomly assigning patients to therapy is the same scientifically as, say, randomly assigning plants to pots, but it is not the same ethically.

Different people view trial results differently, in part for ethical, economic, and political reasons. So there are differences of opinion concerning the need for a confirmatory trial. In this issue of the Journal, Parmar et al. (1) suggest that differences of opinion can be explained from a Bayesian perspective (2) as differences in prior probability distributions of treatment effectiveness. They suggest using a skeptic’s prior distribution to address whether a confirmatory trial is necessary. A confirmatory trial is recommended if, after updating this prior distribution on the basis of results from the original trial, the skeptic’s resulting posterior probability of a clinically important treatment benefit is small. This Bayesian calculation is easy to make, and it can be very useful. Quantifying opinions (of real people as well as of "enthusiasts" and "skeptics") and showing how they are updated on the basis of empirical evidence can be pivotal in a decision process. The point of using posterior probabilities is not to dictate the final decision but to elucidate the decision process. Decision makers are not constrained to choose the option indicated by such an analysis, but if they choose otherwise, they should be able to identify which assumptions in the analysis are wrong, and they can redo the analysis with the repaired assumptions. Anyone considering a confirmatory trial will find the analyses of Parmar et al. instructive. Furthermore, enthusiastic and skeptical posterior probability distributions of treatment benefit would be useful additions to the discussion sections of clinical trial reports.

In addressing the decision to confirm a trial, Parmar et al. restrict their consideration to data from the original trial. This is seldom the only available information. Consider the authors’ example of the Cancer and Leukemia Group B (CALGB) trial of nonresectable stage III non-small-cell lung cancer (3,4). Prior to or during the initiation of the confirmatory trial (5), several other randomized trials (6-9) had compared radiotherapy plus chemotherapy with radiotherapy alone, although none used exactly the same chemotherapy regimen as did the CALGB. This other evidence could be included in a statistical analysis, and the Bayesian approach is ideal for synthesis. However, it would be wrong to regard patients from different studies as exchangeable and simply pool them. Hierarchical models (10-12) provide an appropriate alternative. In a hierarchical analysis, patients and trials are viewed as two different levels of experimental units. Patients are regarded as sampled from a trial’s population, and trials are regarded as sampled from some larger population—a random effects model. The original trial is a sample of size 1 from this population, and this is so irrespective of the trial’s size.

Another virtue of a Bayesian approach is that it fits very well into a decision analysis. Parmar et al. indicate that a confirmatory trial may not be appropriate, despite a low (skeptical) probability of a clinically meaningful benefit if the disease being treated is rare or if there are no reasonable alternative therapies. More generally, the issues of prevalence of disease and availability of alternative therapies can be neatly incorporated into a formal decision analysis, a topic beyond the scope of this editorial (13,14).

Despite its advantages, a Bayesian analysis that ignores the manner of selection of clinical trials as candidates for confirmation is no better than any other analysis. Trials that are chosen

*Correspondence to: Donald A. Berry, Ph.D., Institute of Statistics and Decision Sciences, Duke University, 223 Old Chemistry Bldg., Durham, NC 27708-0251.
for replication tend to be those showing the greatest therapeutic advantage. Observed differences in therapies have at least two possible explanations: they are real and they are the result of random variation. Whether or not the first applies, the second almost always does. A trial selected because of a large observed therapeutic difference is biased. Replicating such a trial will likely give a smaller difference, even if the circumstances of the two trials are identical.

This bias results from the regression effect—also called regression to the mean. It is ubiquitous and insidious. Failure to appreciate its presence is among the most serious of errors made in medical (and other) research (2). It is possible to adjust for biases resulting from the regression effect without a confirmatory trial—thecoretically at least. One method is to use a hierarchical model including all clinical trials. Appropriate adjustments depend on many things, including sample sizes of the various trials. The spirit of such an adjustment is to convey: Estimating the true benefit of a therapy to be a fraction—such as one half—of the observed benefit is likely to be closer than the observed benefit itself. (Baseball sophisticates know that a player’s batting average in the current year is not a good estimate of his batting average next year and that a better estimate is the simple average of his batting average this year and this year’s batting average of the entire league.) However, the most appropriate adjustment is not obvious because trials are often considered for confirmation for reasons other than observed treatment differences. Other reasons include the following: medical importance, the presence of corroboratory clinical or preclinical information, economics, and politics. In a particular trial, it is difficult to know how much of an observed treatment benefit is real and how much is random. A subjective assessment using a Bayesian approach is possible (2). However, this may vary considerably from one assessor to another and not lead to a consensus.

Overcoming the regression effect is itself a good reason to replicate a trial. Indeed, since large deviations are usually the ones checked by replication and the replicated results are usually smaller, the regression effect has probably played an important role in establishing replication as a scientific standard. In cases where the second trial shows less of a treatment difference than the first—the typical case, and the case in examples considered by Parmar et al.—one might average the results of the two trials to obtain an overall estimate. However, because of selection and regression-effect bias, the second trial’s results have more credibility and should be given more weight. In extreme cases, the first trial might reasonably be discounted entirely.

There are settings in which the decision to carry out a confirmatory trial is not difficult. One setting is when the results of the original trial have, at most, a moderate impact on clinical practice. Apparently, such was the case after the CALGB lung cancer trial that compared radiotherapy plus chemotherapy with radiotherapy alone (15). Moreover, one of the confirmatory trial’s authors (16) suggests that, even now, patients are usually treated with thoracic irradiation alone.

Parmar et al. do not address the closely related question of monitoring trials with the possibility of early stopping. The circumstances in the previous paragraph apply as well when considering stopping a trial prematurely. If a trial’s monitoring committee can foresee that, despite publication of the trial’s results, most patients in clinical practice would continue to be treated with the arm that is the apparent loser—radiotherapy alone, for example—then it is not ethically imperative to stop the trial. On the other hand, it may be reasonable to stop an apparently positive trial early simply to make way for a confirmatory trial.

To summarize, confirmatory trials are always important from a scientific perspective. Whether they are ethical is another, much more complicated matter. The methods of Parmar et al. are useful in assessing the need to confirm a trial, with two reservations: 1) The setting of the original trial may be special and Bayesian or other analyses should recognize the possibility of trial differences, and 2) the regression effect makes it difficult to assess the magnitude of a treatment benefit on the basis of information from any single trial.

References


References