Negative Results of Randomized Clinical Trials Published in the Surgical Literature

Equivalency or Error?

Justin B. Dimick, MD; Marie Diener-West, PhD; Pamela A. Lipsett, MD

Hypothesis: We hypothesized that review of randomized controlled clinical trials (RCTs) with nonstatistically significant or “negative” results published in the surgical literature do not have appropriate statistical power to demonstrate equivalency between treatment arms.

Data Sources and Study Selection: The MEDLINE database was searched to obtain reports of all RCTs with negative results published in 3 surgical journals from 1988 to 1998. Manual review of one year (1997) of publications for each journal was performed to validate our search strategy. Equivalency was evaluated using the Two One-Sided Tests Procedure and post hoc power calculations.

Data Synthesis: Ninety reports of RCTs with negative results were identified in the surgical literature between 1988 and 1998. The manual review of 1997 showed a 100% retrieval rate for our search strategy. After applying the Two One-Sided Tests Procedure, 35 reports (39%) met the criteria for demonstrating equivalency. The other 55 reports (61%) contained at least a 10% absolute difference in the 90% confidence interval of . Using the power calculation method, only 22 (24%) articles had a power greater than .80 to detect a 50% difference in therapeutic effect. Only 29% of the reports included a formal sample size calculation and these studies were more likely to demonstrate equivalency than those without a sample size estimate (P<.01).

Conclusions: Many reports from negative RCTs published in the surgical literature lack sufficient statistical power to establish that clinically important differences are not present. Surgeons should perform appropriate sample size calculations when designing RCTs and recognize the utility of confidence intervals when reporting negative results.

Arch Surg. 2001;136:796-800

Clinical decisions should be based on the critical appraisal of relevant literature coupled with the experience and judgment of the surgeon. The randomized controlled clinical trial (RCT) is the definitive method to investigate the relative efficacy of 2 or more interventions of interest. However, RCTs comprise only 3% to 7% of research publications in surgical journals. Previous efforts aimed at evaluating the quality of surgical RCTs have shown that many of them contain errors in methodology.

When reporting the results of a clinical trial, investigators often state whether the results of a comparison between treatment groups demonstrate a statistically significant difference with respect to the primary outcome or end point. This statement refers to the P value, obtained after applying a statistical hypothesis test. If P is less than some predefined probability (usually .05), the 2 groups are considered statistically different. When P is greater than .05, it is concluded that differences between the groups may be explained by chance alone.

However, there are 2 types of errors owing to chance that may result during statistical hypothesis testing (Table 1). A type I error concludes, based on the P value, that there is a difference between the intervention and nonintervention groups when one does not exist. A type II error concludes that there is not a difference between the treatment groups when one may exist. The power of a study (β) is the probability of a statistically important difference between the 2 groups when such a difference exists. Therefore, in a trial in which the 2 therapeutic options seem the same, the underlying statistical power of the study to detect a true difference between the groups must be considered. Reporting P>.05 is not the same as demonstrating equivalency between 2 treatment options.

Because surgical interventions are complex and patient or physician preference may limit enrollment, the sample sizes of many RCTs in the field of surgery are small. Consequently, the trials may have inadequate statistical power to detect clinically important differences in therapeutic effects. This study was undertaken to estimate the prevalence of studies at risk for type II errors in the surgical specialty.
METHODS

LITERATURE SEARCH

The search strategy was designed to yield a sample of RCTs published in the surgical specialty literature from which “negative” trials (those that concluded that there was no difference between the treatment arms) could be selected. The MEDLINE database was searched using the Medical Subject Headings (MeSH): clinical trials and RCTs, the keywords clinical trials and RCTs, and the publication types RCTs and controlled clinical trials. The search was limited to 3 surgery specialty journals (Annals of Surgery, Surgery, and Archives of Surgery) from January 1988 to December 1998.

The abstracts from all articles were reviewed; our analysis included all reports of RCTs that concluded there were equivalent dichotomous outcomes in the treatment arms. The statement regarding equivalency had to be explicit (for example, “there was no statistically significant difference between the groups”) and had to refer to a statistical test with $P > 0.05$ for the outcome variable of interest. The outcome variable we chose was either clearly labeled as the primary end point or was the primary focus of the article. Articles were excluded for the following reasons: not representing original data (eg, meta-analysis, review articles, editorials, letters); nonrandomized treatment allocation; the use of animal subjects; retrospective data collection; and having a continuous variable as the primary end point. To document the adequacy of our literature search, a manual review of one year (1997) of publications of each journal was performed. Using this information we calculated the percent yield of our MEDLINE search strategy.

DATA ABSTRACTION

The full text of the included articles was systematically reviewed. Data were abstracted and recorded on a standardized form. Information was recorded regarding the type of intervention (surgical, pharmacological, adjuvant oncologic therapy, or other); author affiliations (surgery, anesthesia, medicine, biostatistics, or other); the presence of an a priori power calculation; the event rates in the 2 treatment arms; the number of subjects in each treatment arm; the presence of a post hoc power calculation; and the discussion of lack of power as a weakness. There was a single data abstractor who was responsible for primary review of the full text articles (J.B.D.). To assess the accuracy of our data abstraction, 20 articles were randomly chosen and a second author (P.A.L.) reviewed the full text, repeating the data abstraction. The percent agreement and the $k$ statistic (the percent agreement greater than that expected by chance alone) were calculated to assess interobserver variability for each outcome variable. Each study received a number and data analysis was blinded to the author and institution of the publication.

CONFIDENCE INTERVALS (CIs) AND EQUIVALENCY TESTING

The goal of an experiment evaluating a new therapy is to estimate the proportion of patients who achieve the outcome in the treatment group (PT) and in the control group (PC). $\Delta$ is the difference between the 2 groups. Because the entire population of similar patients is not studied, it was necessary to assess the precision with which our estimate was likely to represent the true difference between the groups. This was accomplished by providing a range of values based on the observed data that were consistent with the true value. The precision of the observed difference ($\Delta$) between the treatment groups was best represented by a CI for the true population difference. To calculate the 90% CI for $\Delta$, the standard error ($SE$) and upper and lower limits of a 90% CI were calculated using the sample size of the control (n1) and treatment (n2) groups and the following equations:

$$SE = \frac{PC(1-PC)}{n1} + \frac{PT(1-PT)}{n2}$$

Upper limit 90% CI = $\Delta + 1.645 \times SE$

Lower limit 90% CI = $\Delta - 1.645 \times SE$

To determine which reports demonstrated true negative results (equivalency) we used the Two One-Sided Tests Procedure. First, a $(1 - 2 \times \alpha)$ CI was constructed for the absolute difference between treatment groups for each study. Equivalency was concluded if the limits of the CI were entirely within a predetermined equivalency interval. For most studies, $\alpha$ was set at .05 and 90% CIs were calculated. We considered plus or minus 10% and 25% absolute differences to be clinically important. For instance, if the event rate was 10% in the experimental group and 20% in the control group, the absolute difference between the 2 groups ($\Delta$) would be 10%, or a relative reduction of 50%. As is commonly the case in the literature, we considered these differences to be clinically important.

POWER CALCULATIONS

The power to detect a predefined effect size was calculated for each RCT. We chose to calculate the power needed to detect a difference ($\Delta$) of 25% and 50% between the 2 groups given the baseline event rate. For each publication, the event rate in the PC group was determined and the proportion that represented a 25% [PC – (PC) (0.25)] and a 50% [PC – (PC) (0.5)] reduction were calculated. Using the number of patients in each treatment arm and the proportions representing a 25% and 50% reduction in the baseline event rate, the power of each study was calculated setting the $\alpha$ at .05 (2-tailed). Publications reporting an event rate of 0 in both treatment groups were excluded from the power analysis but were included in the assessment of other end points.

OUTCOME VARIABLES

The primary outcome of our investigation was whether the reported results of an RCT met the criteria for equivalency using the Two One-Sided Tests Procedure. One secondary outcome was the post hoc power calculation for each report. An unacceptable level of a type II error was defined as any post hoc power less than 80% ($P > .2$). Other secondary outcomes of this study were the presence of an a priori power calculation, presence of a post hoc power calculation, and discussion of lack of power as a weakness of the study. In addition, the reports were divided according to type of intervention (surgical procedure; adjuvant therapy [chemotherapy, external beam or intraoperative...
literature and to discuss the implications of our findings on study design, reporting, and interpretation of RCTs.

### RESULTS

The MEDLINE search strategy yielded 526 publications (Figure 1). After applying the initial exclusion criteria, 268 prospective RCTs using human subjects remained. Randomized clinical trials represented 3.2% of the total number of publications during the 11-year study period (1988-1998). These studies included reviews, case reports, and clinical and basic science articles. Therefore, the rate of clinical trials that were randomized is likely higher than this number.

Of these abstracts, 136 (51%) had conclusions stating that the outcomes were equivalent in the treatment arms. Full text was obtained for further review and 8 additional articles were excluded because they failed to explicitly conclude equivalency between the treatment groups within the text. These 8 abstracts stated that the 2 treatments being compared may be equivalent but the body of the article suggested that there may be no difference between the outcomes, especially when more than 1 outcome is considered and the primary outcome is not evident. In addition, 32 articles that had a continuous rather than a dichotomous variable as the primary outcome were excluded. The remaining 96 RCTs reported dichotomous primary outcomes that explicitly concluded equivalency between the treatment groups. The manual search of the target journals for 1997 demonstrated that our MEDLINE search strategy retrieved 26 (100%) RCTs published during that year.

Table 2 presents the percent agreement and $\kappa$ statistics associated with interobserver variability in data abstraction for the 20 randomly selected articles. There was excellent agreement between observers in assessing the risk of a type II error ($\kappa = 1.0$ [100% agreement]) and the presence of an a priori sample size calculation ($\kappa = 0.89$ [93% agreement]). Both of these values were interpreted as having “almost perfect” agreement. The more subjective assessments, such as the presence of post hoc power calculation ($\kappa = 0.59$; 85% agreement) and discussion of lack of power as a limitation ($\kappa = 0.61$; 80% agreement), demonstrated more interobserver variability; these values were interpreted as having “moderate” and “substantial” agreement, respectively.

Table 3 presents the journals of publication and associated characteristics of the 96 articles included in the analysis. There seemed to be no increase in publication of RCTs during the 11-year period, with approximately equal numbers derived from each time interval. Most trials tested the efficacy of either a surgical procedure (n=43) or a pharmacological agent (n=42) and the remainder involved adjuvant cancer therapy (n=6) or other (n=5). Surgeons were the sole authors in 50 articles (52%) and they shared authorship predominantly with colleagues in the departments of medicine (n=14), anesthesia (n=8), and biostatistics (n=6).

The Two-Sided Tests Procedure was performed on 90 articles. Six articles were not included in the equivalence analysis because the event rate was 0 in both of the treatment groups. Of the included articles, 35 (39%) demonstrated equivalency (given an equivalence interval of ±10% absolute difference). The 90% CIs for the differences between treatment and control groups are shown in Figure 2. In the power analysis, none of the articles demonstrated 80% or greater power to detect a 25% relative difference in the treatment groups and only 24% had a power of 80% or more to detect a 50% relative difference. Of the reports of RCTs that were at risk for type II errors, only 14 (19%) mentioned a small sample size or lack of power as a weakness. Furthermore, only 7 articles (9%) presented a post hoc power analysis, formally addressing the lack of
power in their study. Twenty-eight trials (29%) included an explicit sample size calculation in the report and these trials were less likely to be at risk for a type II error (P < .01).

This study documents that results of many reports of negative clinical trials published in the surgical literature lack precision or are at risk for a type II error and do not demonstrate equivalency. In other words, the reports may conclude there is no difference between intervention and control or placebo groups when one may exist. We used 2 approaches to assess the risk of a false conclusion and both demonstrated similar results. Specifically, using an estimation approach for equivalency testing, only 39% of reports satisfied the criteria for equivalency. Likewise, using a hypothesis testing approach, only 24% of the articles had a power greater than 80% to detect a 50% difference between the treatment arms. Thus, 61% of the RCTs were failed experiments in that the researchers failed to reach a conclusion, either of equivalency or dissimilarity. Such failures can be prevented by appropriate a priori power considerations. Furthermore, it was shown that many authors do not include a formal sample size calculation or discuss lack of power as a limitation. These findings have important implications on surgical decision-making. If studies with inadequate statistical power or sample size fail to demonstrate benefit of a particular therapy, we may inappropriately label that therapy as ineffective and refrain from pursuing further research, effectively abandoning a potentially efficacious therapy.

In a landmark study published in 1978, Freiman et al. conducted a survey of 71 negative RCTs published in the medical literature. They demonstrated that reports of many RCTs do not have adequate power to demonstrate clinically important differences in therapeutic effect. Of the 71 trials evaluated, 67 had less than a 90% power to detect a 25% therapeutic improvement and 50 had less than a 90% power to detect a 50% improvement. We chose to assess an 80% rather than a 90% had less than a 90% power to detect a 50% therapeutic improvement and 50 effect. Of the 71 trials evaluated, 67 had less than a 90% power to demonstrate clinically important differences in therapeutic samples of many RCTs do not have adequate power to demonstrate equivalency. In other words, the reports may conclude there is no difference between intervention and control or placebo groups when one may exist. We used 2 approaches to assess the risk of a false conclusion and both demonstrated similar results. Specifically, using an estimation approach for equivalency testing, only 39% of reports satisfied the criteria for equivalency. Likewise, using a hypothesis testing approach, only 24% of the articles had a power greater than 80% to detect a 50% difference between the treatment arms. Thus, 61% of the RCTs were failed experiments in that the researchers failed to reach a conclusion, either of equivalency or dissimilarity. Such failures can be prevented by appropriate a priori power considerations. Furthermore, it was shown that many authors do not include a formal sample size calculation or discuss lack of power as a limitation. These findings have important implications on surgical decision-making. If studies with inadequate statistical power or sample size fail to demonstrate benefit of a particular therapy, we may inappropriately label that therapy as ineffective and refrain from pursuing further research, effectively abandoning a potentially efficacious therapy.

In a landmark study published in 1978, Freiman et al. conducted a survey of 71 negative RCTs published in the medical literature. They demonstrated that reports of many RCTs do not have adequate power to demonstrate clinically important differences in therapeutic effect. Of the 71 trials evaluated, 67 had less than a 90% power to detect a 25% therapeutic improvement and 50 had less than a 90% power to detect a 50% improvement. We chose to assess an 80% rather than a 90% power to detect a 50% difference between the treatment arms. Thus, 61% of the RCTs were failed experiments in that the researchers failed to reach a conclusion, either of equivalency or dissimilarity. Such failures can be prevented by appropriate a priori power considerations. Furthermore, it was shown that many authors do not include a formal sample size calculation or discuss lack of power as a limitation. These findings have important implications on surgical decision-making. If studies with inadequate statistical power or sample size fail to demonstrate benefit of a particular therapy, we may inappropriately label that therapy as ineffective and refrain from pursuing further research, effectively abandoning a potentially efficacious therapy.

Table 3. Journal of Publication and Characteristics of Reports From Selected Randomized Clinical Trials

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. (%) of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal</td>
<td></td>
</tr>
<tr>
<td>Annals of Surgery</td>
<td>35 (36)</td>
</tr>
<tr>
<td>Archives of Surgery</td>
<td>40 (42)</td>
</tr>
<tr>
<td>Surgery</td>
<td>21 (22)</td>
</tr>
<tr>
<td>Year published</td>
<td></td>
</tr>
<tr>
<td>1988-1989</td>
<td>20 (21)</td>
</tr>
<tr>
<td>1990-1991</td>
<td>18 (19)</td>
</tr>
<tr>
<td>1994-1995</td>
<td>15 (16)</td>
</tr>
<tr>
<td>1996-1997</td>
<td>19 (20)</td>
</tr>
<tr>
<td>1998</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Author affiliation</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>50 (52)</td>
</tr>
<tr>
<td>Surgery and medicine</td>
<td>14 (15)</td>
</tr>
<tr>
<td>Surgery and anesthesia</td>
<td>8 (8)</td>
</tr>
<tr>
<td>Surgery and biostatistics</td>
<td>6 (6)</td>
</tr>
<tr>
<td>Surgery and other</td>
<td>16 (17)</td>
</tr>
<tr>
<td>Nonsurgical</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Intervention</td>
<td></td>
</tr>
<tr>
<td>Surgical procedure</td>
<td>43 (45)</td>
</tr>
<tr>
<td>Pharmacological agent</td>
<td>42 (44)</td>
</tr>
<tr>
<td>Adjuvant therapy</td>
<td>6 (6)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Total</td>
<td>96 (100)</td>
</tr>
</tbody>
</table>

Figure 2. Point estimates of the difference between treatment groups (Δ) with 90% confidence intervals for 90 negative randomized clinical trials. (α) is determined by the investigator and is often set at a level of 5% or less. Unlike the type I error, the probability of a type II error (β) is not set by the investigator; it is a function of α, the size of the study population, the frequency of the outcome of interest, and the magnitude of the difference between the 2 groups being studied (Δ). For our study, we choose an absolute difference of 10% and 25% (Two One-Sided Tests Procedure) and a relative difference of 50% for the power calculations because the magnitude of these differences is commonly considered to be clinically significant.

A type II error relies, in part, on the results of the RCT and is therefore subject to unplanned variation. Furthermore, the type II error is not commonly addressed in the discussions of many reports of RCTs and may be an overlooked source of inaccurate interpretation of clinical trial...
results. There may be great consequence in concluding that 2 therapeutic options are the same when, in fact, they are clinically significantly different. Our study demonstrates that most RCTs published in surgical specialty journals are at risk for this type of error, assuming that the magnitude of difference between the treatment options is 50%. When designing clinical trials in the future, surgeons should calculate both a priori sample size and power in the early planning stages. Obtaining an estimate of the required sample size early in planning may help refine the research question. For instance, if the sample is prohibitively large, choosing a surrogate outcome variable with a higher frequency may allow a smaller sample. On the contrary, the trial may be judged as unfeasible before significant resources are needlessly spent. Only 29% of the reports of RCTs evaluated in our study included sample size estimates, but those doing so were less likely to be at risk for a type II error. Solomon et al\textsuperscript{10} report a similarly low percentage of reports of RCTs in surgery with adequate sample size determinations (19% and 11%, respectively).

Performing a sample size calculation is relatively straightforward and can be conducted using any statistical software package or tables available in most statistical texts.\textsuperscript{17} First, the expected frequency of the outcome variable in the nonintervention group must be estimated; this is usually taken from previously published research or, in some instances, from a pilot study. Second, the smallest difference that is felt to be clinically meaningful (\(\Delta\)) between the 2 treatment arms must be chosen. Usually, a relative change of 25% to 50% from the baseline event rate is generally accepted. Third, the minimally acceptable probability of a type I or type II statistical error must be chosen. The generally accepted value of \(\beta\) is .20 or .10 (power of .80 to .90). However, in a trial specifically designed to demonstrate equivalency of 2 therapeutic options (in which a negative result is anticipated) a smaller \(\beta\) (or larger power) is desired since avoiding a type II error in this setting is particularly important.\textsuperscript{5,6} In our study, we set the power at .80, which is the minimally acceptable level, especially for trials that claim equivalency.

When surgeons obtain a negative result they should report the associated 90% or 95% CI of the difference in outcome between the groups (\(\Delta\)). This is necessary to appropriately interpret the results and perceive the risk of a type II error. If the CI contains clinically significant differences in the outcome of interest, the use of the Two Sided Hypothesis Procedure for demonstrating equivalency will assist in understanding the differences between 2 treatment options. In this technique, the precision of the absolute difference (\(\Delta\)) in the primary outcome variable between treatment groups is tested by creating a 90% CI. Before the calculation is obtained we had to estimate what we consider a clinically significant difference between the groups. In general, a 10% absolute difference (eg, 10% in group A and 20% in group B—a 50% relative difference) would be clinically significant. If this predefined difference is included within the 90% CI we can be relatively sure that the treatment options are equivalent. Otherwise, the trial fails to demonstrate equivalency and we have not given the intervention of interest a fair trial.

Our study has several limitations. We did not include all surgical research published during the represen-

tative period. Some RCTs are not published in the surgical literature, but appear in larger multidisciplinary journals. However, this study was designed to examine the surgical specialty literature and calculate the frequency of type II errors in those journals. Also, we used a search of a computerized database (MEDLINE) to locate articles running the risk of missing a certain proportion of RCTs. Previous authors have shown MEDLINE to retrieve less than half of RCTs published during a given period.\textsuperscript{1} To minimize this risk, we performed a manual search of the target journals for a 1-year period; this effort demonstrated a 100% retrieval rate for our search strategy, allowing us to be confident that the articles included represented most of the RCTs during the study period. An additional limitation comes from the assumptions made in performing the power calculations. We assigned the same \(\Delta\) to each trial (50%).

These findings have important implications on the future design and interpretation of RCTs in surgery. Surgeons should conduct a sample size calculation during the design phase of the study. In addition, they should report the CI of the difference between the treatment groups. If there is inadequate statistical power to detect clinically significant differences between treatment groups, this should be explicitly stated in the conclusion. Such practice may better inform the reader and promote further study of potentially efficacious therapies.

Corresponding author and reprints: Pamela A. Lipsett, MD, Department of Surgery, Johns Hopkins Hospital, 600 N Wolfe St, Baltimore, MD 21287-4683 (e-mail: plipsett@jhmi.edu).

REFERENCES