

Claims of Equivalence in Medical Research: Are They Supported by the Evidence?

William L. Greene, MD; John Concato, MD, MPH; and Alvan R. Feinstein, MD

Background: Most clinical studies are done to show comparative superiority, but many reports now claim equivalence between the investigated entities. These assertions may not always be supported by the methods used and the results obtained.

Purpose: To assess the justification and support for claims of clinical or therapeutic equivalence in medical journals.

Data Sources: A search of MEDLINE for articles published from 1992 through 1996.

Study Selection: From 1209 citations that contained the word *equivalence* in the title or abstract or contained the Medical Subject Heading *therapeutic equivalency*, we excluded 1121 studies reporting nonoriginal research, purely laboratory or other nonhuman research, and studies in which equivalence was not the main claim. The remaining 88 eligible papers were evaluated for five methodologic attributes.

Data Synthesis: Only 45 (51%) of the 88 reports were specifically aimed at studying equivalence; the others either tried to show superiority or did not state a research aim. The quantitative distinctions regarded as "equivalent" ranged from 0% to 21% for direct increments and from 0% to 76% for proportionate differences. An equivalence boundary was set and confirmed with an appropriate statistical test in only 23% of reports. In 67% of reports, equivalence was declared after a failed test for comparative superiority, and in 10%, the claim of equivalence was not statistically evaluated. The sample size needed to confirm results had been calculated in advance for only 33% of reports. Sample size was 20 patients per group or fewer in 25% of reports.

Conclusions: Many studies of clinical equivalence do not set boundaries for equivalence. Claims of "difference" or "similarity" are often made not by thoughtful examination of the data but by tests of statistical significance that are often misapplied or accompanied by inadequate sample sizes. These methodologic flaws can lead to false claims, inconsistencies, and harm to patients.

Most clinical research activities are aimed at showing that one agent or method is better than another. An increasing number of reports, however, now conclude that the investigated entities are "equivalent." Thus, a new drug or treatment may be deemed just as effective as a standard therapy while being, for example, less costly or easier to use (1). Equivalence can also be claimed for generic versions of innovator drugs (2) and for such diverse entities as medical protocols (3), surgical techniques (4), and medical devices (5). As physicians, insurers, and hospitals put increasing emphasis on practicing evidence-based medicine, claims of substantial treatment benefit have come under scrutiny. In contrast, claims of therapeutic equivalence may not be reviewed with the same quantitative rigor. This can lead to patient harm if clinically inferior treatments are erroneously deemed equivalent to a standard approach or if potentially superior therapies are discarded as merely "equivalent."

Despite the many scientific and statistical procedures used to confirm that a large difference is "significant," less attention has been given to the logic and methods for establishing equivalence (6). In the context of hypothesis testing, equivalence exists only as a theoretical entity—an infinitely large sample size would be needed to unequivocally establish no difference between compared groups. In practice, an observed difference can be compared to a specified value considered "small" (that is, not clinically important). Unfortunately, there are no established "gold standard" criteria for how to construct and support such an equivalence claim. Proposed approaches include using confidence intervals to exclude "clinically meaningful differences" (7) or applying variations of the analytic strategy of rejecting the null hypothesis (8).

In this context, testing for equivalence involves rejecting an "alternative" hypothesis of a large difference between examined groups or entities. Yet, investigations that report clinical equivalence may not use this approach. Instead, after finding a negative result in conventional tests for statistical significance (for example, $P > 0.05$), investigators may declare that the entities compared are "equivalent."

Our current study was done to determine

Ann Intern Med. 2000;132:715-722.

For author affiliations and current addresses, see end of text.

whether published claims of equivalence are supported by the methods used and the results obtained.

Methods

Study Sample

Using the National Library of Medicine Medical Subject Heading (MeSH) term *therapeutic equivalency* and the text word *equivalence*, we did a structured MEDLINE search for English-language original research reports published from 1992 through 1996 in which equivalence was claimed. The MeSH term searching process identifies papers in which the National Library of Medicine has identified a main point in the text. The separate text word searching process indicates papers in which the selected word is used in the title or abstract. We used both of these search strategies to obtain a representative, but not exhaustive, sample of published reports claiming equivalence. Reports identified with the specific MeSH term *therapeutic equivalency* are more likely to be methodologically sound, and papers defined by the text word *equivalence* are more likely to state the aim of the study than if they had been selected from a more general word, such as *equivalent* or *equal*.

We reviewed the obtained citations on-line to select, for further evaluation, those that appeared likely to describe original research claiming equivalence. We included randomized clinical trials as well as pertinent observational studies (for example, reference 9) because although these types of research are conducted by using different methods, we believe that the quantitative claim of equivalence can be substantiated similarly.

After reviewing the title and, if necessary, the abstract, we discarded citations to papers, such as reviews, meta-analyses, or commentaries (for example, references 10 and 11) that did not report original data. We also excluded reports (for example, references 12 and 13) of purely laboratory or other nonhuman research, as well as those intended solely to show pharmacokinetic bioequivalence, such as generic drug applications to the U.S. Food and Drug Administration. Finally, we excluded papers that reported equivalence for anything other than patients' outcomes, such as a comparison of two radiation therapy regimens in which "dose equivalence" referred only to standardized radiation fractions (14).

For the remaining potentially eligible citations, abstracts were further reviewed according to the preceding criteria. Whenever the suitability of a paper was uncertain, the entire text was reviewed. All

reviews were done in a structured fashion by one author using a database (Microsoft Access for Office 97, Microsoft, Inc., Redmond, Washington) created for this study. Although rules for inclusion and categorization were defined a priori, decisions were not straightforward for approximately 10% of reports. These difficult papers were reviewed by all three investigators for a consensus decision.

Evaluation of Papers

The entire text of each included paper was evaluated in a structured fashion for five prespecified attributes relating to the assessment and ultimate claim of equivalency. The attributes evaluated are listed below, along with the justification for their choice.

1. Statement of research aim. A stated research goal is needed to allow the investigators to choose the pertinent variables for study, boundaries for the magnitude of an equivalent result, and cogent analytic methods.

2. Magnitude of reported differences. Evaluating the clinical sensibility of an equivalence claim requires knowledge of exactly what is being called a "negligibly small" difference. We therefore tabulated the actual values of the quantities that constituted the difference between the investigated groups. Although these quantities ranged broadly in units of measurement and extent of variability, differences can be standardized as an "effect size" (15), which is usually calculated as the direct increment between groups divided by the standard deviation in the control group. For example, if drug X is successful in $60\% \pm 30\%$ (SD) of patients and placebo is successful in $40\% \pm 20\%$ (SD), then the effect size is $(60\% - 40\%)/20\% = 1.0$. The effect size can thus be considered a unit-free ratio of "signal" to "noise."

3. Choice of quantitative boundary. We determined whether the investigators had set an a priori quantitative boundary for what would constitute equivalence in the reports examined. Demarcating a maximum value of "small," beyond which a difference could no longer be deemed "equivalent," is needed for investigators and readers to appraise the numerical results in a clinical-scientific as well as a statistical manner. The boundary could be set from an absolute difference in means or medians, a proportionate difference in results, a ratio of two-group results, or an "effect size." Although a single criterion does not exist for establishing what is "large," proportionate differences of 20% or greater between clinical groups have been suggested as potentially important (16). In addition, an effect size less than 0.20 has often been considered trivially small, 0.50 has been considered moderate, and 0.80 has been considered large (17).

We are not advocating general use of these arbitrary thresholds, and neither the incremental difference between compared groups, the proportional difference, nor the effect size was used as a “gold standard” criterion for what would constitute an acceptable study for equivalence. These thresholds can, however, serve to describe the magnitude of observed differences in our investigation, and this information is included for illustrative purposes.

4. Method of statistical (stochastic) testing. We next determined what, if any, testing was done to support the claim of equivalence, and we specifically checked whether the claim of equivalence was tested directly or was supported only by a failed test for superiority. In a direct test, the differences observed between groups or patients are compared against a specific equivalence boundary. A direct test of equivalency is intended to reject the “alternative” hypothesis that the true difference is larger than the boundary limit, whereas statistical tests for superiority are aimed at rejecting a “null” hypothesis of no difference.

5. Calculation of sample size. We expected that information about sample sizes, including advance calculations, would be reported. Such information would help, for example, to explain a paradox in which a large observed difference is deemed equivalent because the sample size was too small to achieve statistical significance. This problem of inadequate statistical power (which may cause studies to “miss” important observed differences) has been highlighted previously (18).

“Quality” or “Impact” of Journals

Although the quality of the journal might be considered when methodologic rigor is evaluated for published research, we did not try to define a “high-quality” medical journal. Instead, we determined which papers came from the 119 Abridged Index Medicus (AIM) journals that the National Library of Medicine regards as “selected biomedical journal literature of immediate interest to the practicing physician” (19). The National Library of Medicine states that journals are included in AIM according to “the quality of the journal, usefulness of journal content for the professional, and the need for providing coverage in the fields of clinical medicine” (19). The list includes *Annals of Internal Medicine*, *BMJ*, *JAMA*, *The Lancet*, *The New England Journal of Medicine*, and many leading subspecialty journals but omits such “nonclinical” journals as *Science*, *Nature*, and *Cell*.

Role of the Funding Source

This study was funded by the Robert Wood Johnson Foundation through its Clinical Scholars Program. The funding source had no role in the

Table 1. Evaluation Using Methodologic Criteria

Methodologic Attribute	Studies Adhering to Attribute (n = 88), n (%)
Statement of research aim	50 (57)
Aim of equivalence	45 (51)
Proportionate difference between entities being compared <20%	51 (58)
Quantitative boundary chosen	20 (23)
Method of stochastic testing	
Not done	9 (10)
Failed test for superiority	59 (67)
Tested for equivalence	20 (23)
Required sample size calculated	29 (33)

collection, analysis, or interpretation of the data or in the decision to submit the paper for publication. No proprietary interest exists in relation to this research.

Data Synthesis

Assembly of the Study Sample

The literature search for 1992 through 1996 yielded 1209 citations whose titles were reviewed on-line for potential inclusion; 907 papers were excluded because they described nonoriginal research, reported laboratory or other nonhuman research, or did not refer to equivalence of outcomes. For the remaining 302 citations, the abstracts and, when necessary, the entire text were reviewed according to the described criteria. Among the papers excluded at this step, 57 did not describe original research, 48 reported laboratory or other nonhuman research, 106 either did not claim equivalence or claimed it only for entities other than outcomes, and 3 could not be obtained. The remaining 88 papers constituted the study sample.

Evaluation Using Methodologic Criteria

The results obtained when the 88 papers were reviewed according to our methodologic criteria are discussed in the next five subsections and are summarized in **Table 1**.

Statement of Research Aim

Among the 88 reports, a specific research goal was stated a priori to show equivalence in 38 reports (43%), superiority of one treatment over another in 5 (6%), and both efficacy (compared with placebo) and equivalence in 7 (8%). These results indicate that 50 (57%) reports had a stated research aim and 45 (51%) mentioned equivalence. In the remaining 38 reports (43%), a research goal of superiority or equivalence was not mentioned. In one example of an unspecified goal, two metered-dose inhalers for asthma were studied, without any men-

tion of superiority or equivalence, “to compare the bronchodilator effects of the dry-powder form of Berodual . . . with MDI [metered-dose inhaler] when delivered in an equal dosis [sic]” (20).

Magnitude of Reported Differences

The absolute differences between two groups ranged from 0% (in 7 papers) to 30% (21), and proportionate differences ranged from 0% to 137%, with an average of 16%. In 42% (37 of 88) of reports, the proportionate difference was 20% for at least one of the “equivalent” entities. Examples of these potentially important differences include clinical cure rates of 73% compared with 56% (24% proportionate difference) for two antibiotics used to treat lower respiratory tract infections (22), and median survival of 25 months compared with 44 months (proportionate difference, 76%) for two different treatments for laryngeal cancer (4).

Although we used the authors’ data to calculate the effect size wherever possible, only 35% (31 of 88) of papers in the study provided enough information for the calculation. The other 57 offered no information about the standard deviation or the standard error of the mean. The effect size, when available, was a “small” value of less than 0.20 in 20 reports, a “moderate” effect of 0.20 to 0.79 in 7 reports, and a “large” effect of 0.8 in 4 reports; in 1 report (23), the effect size was a very large value of 20.

Choice of Quantitative Boundary

A quantitative boundary was set for equivalence in 20 (23%) papers, but the size and construction of these boundaries varied substantially. The limits for absolute differences ranged from as small as a 1.5% incremental difference in deep venous thrombosis rates (24) or a 1% incremental difference in mortality rates (25), to a 50% incremental difference in seizure rates for different versions of carbamazepine (2). The boundaries for proportionate differences ranged from as low as a 10% relative difference in mortality to as high as a 400% relative difference in “bronchial responsiveness” with two different inhalers (5). In other reports, the criterion for equivalence was an overlap of confidence intervals around mean values. The required degree of overlap ranged from 10% on either end of the confidence interval (26) to a statement that equivalence occurred “if the 95% confidence intervals overlapped” (27).

Method of Statistical (Stochastic) Testing

In 9 reports (10%), the authors declared equivalence after neither setting a boundary for equivalence nor using any formal stochastic tests. In 59 reports (67%), the authors set no equivalence boundary but claimed equivalence after a standard test for superiority (such as a *t*-test or chi-square

test) returned a “nonsignificant” result, that is, a *P* value greater than 0.05.

Only 20 (23%) reports set an equivalence boundary and confirmed it statistically. The tests were all variations of either confidence intervals or a *Z*-type test. The confidence interval strategy, used in 12 papers, establishes a quantitative boundary for “equivalence” in comparisons expressed as a direct increment, a proportional increment, or a ratio of the sample means. After confidence intervals (typically 90% or 95%) are constructed and placed around the sample means (or their increments), equivalence is claimed if the upper limit of the confidence interval is smaller than the boundary. For example, a 1% increment may be chosen as a quantitative boundary. If the observed increment is 0.5% with a 95% confidence interval extending from 0.2% to 0.8%, the compared entities would be deemed equivalent. In a variation of this strategy, a boundary is set for the overlap (such as 10%) in the upper bounds of confidence intervals around sample means. Equivalence is then claimed if the two confidence intervals (for each compared group) do not differ by more than the specified boundary (for example, 10%).

The *Z*-test approach, used in 8 papers (9%), reverses the customary statistical tests for superiority, in which a “big” difference between groups is inferred via rejection of the “null” hypothesis of no difference. In studies of equivalence, a similar strategy evaluates the hypothesis that the observed increment is larger than a limit set for a “small” difference between examined groups. The hypothesis is rejected, and equivalence proclaimed, if the *P* value is sufficiently small (typically <0.05).

Calculation of Sample Size

Only 33% (29 of 88) of reports mentioned an advance calculation of the minimum sample size needed for confirming a quantitatively significant (“large” or “small”) difference, and only 19 of those reports described the method or the assumptions used for the calculations. Of the 29 reports that calculated an advance sample size, only 20 achieved it. The group sizes exceeded 50 patients per group in 41 (47%) of the studies and were 20 per group or fewer in 22 (25%).

Reports Satisfying All Criteria

Overall, 19 reports (22%) satisfied the requirements of stating a research aim of equivalence, defining a quantitative boundary for equivalence, calculating the number of patients needed to statistically demonstrate equivalence, and actually testing statistically for equivalence. An example of a study (24) satisfying the requirements compared subcutaneous low-molecular-weight heparin to standard

heparin for prevention of thromboembolism. The authors state that their aim was to demonstrate equivalence in the clinical outcome of thromboembolism. They decided that a 1.5% absolute increase in outcome events would be meaningful (with an assumed baseline rate of 1%). They calculated that to have 80% power to detect this difference, they would need to enroll 700 patients in each arm of the study. Of 710 patients in the unfractionated heparin group, 4 (0.56%) had an episode of thromboembolism. Of 726 patients in the low-molecular-weight heparin group, 6 (0.83%) had thromboembolism. The observed difference of 0.27% was well below the prespecified equivalence boundary of 1.5%, and the authors declared equivalence.

“Quality” or “Impact” of Journals

Of the 88 papers in the study, 35 were published in AIM journals and 53 were not. As shown in **Table 2**, the reports in the non-AIM journals were slightly more likely to satisfy our criteria than those in the AIM journals. A higher proportion of the non-AIM papers defined equivalence as a goal of the study, established a zone for equivalence, calculated an advance sample size to achieve the desired outcome, and actually tested for equivalence. Six of the 35 AIM papers (17%) satisfied four of the methodologic criteria, compared with 13 of the 53 non-AIM papers (25%). (The fifth criterion, magnitude of reported differences, is a descriptive issue and was not examined in this context.)

Discussion

Studies claiming equivalence make a common assertion: that the effect or difference demonstrated, whatever it is, is not sufficiently large to “matter.” No standard methods or criteria exist for demarcating what is a “large” or an important effect (for example, a 1% incremental difference in mortality between two treatment regimens for myocardial infarction or a proportionate difference of 15% in microbiological cure rates for two antibiotics). We have not critiqued the judgment of the authors of the studies analyzed here in terms of their choice of what is “large” or “small.” Rather, we have focused on the choice of methods for demonstrating the “unimportance” of differences found. We found that more than three quarters of clinical studies in our sample of the medical literature claiming equivalence did not specify that the research is intended to show equivalence, set a quantitative boundary for the magnitude of equivalence, use a sample size adequate to detect a meaningful difference between the groups tested, or support the equivalence claim with an appropriate statistical test.

Table 2. Performance in Abridged Index Medicus Compared with Non-Abridged Index Medicus Journals*

Methodologic Attribute	Papers in AIM Journals (n = 35)	Papers in Non-AIM Journals (n = 53)
	n (%)	
Statement of research aim	17 (49)	33 (63)
Aim of equivalence	15 (43)	30 (57)
Quantitative boundary chosen	6 (17)	14 (26)
Tested for equivalence	6 (17)	13 (36)
Required sample size calculated	10 (29)	19 (36)
Satisfied all four criteria	6 (17)	13 (25)

* AIM = Abridged Index Medicus.

When research results are analyzed and their conclusions supported with statistics, a logical, methodologically sound approach is needed. The issue of how best to demonstrate the “unimportance” of an observed difference has been discussed by statisticians, epidemiologists, and methodologists. In 1977, Dunnett and Gent (28) laid out the mathematical rationale for specifying an “alternative” hypothesis of a meaningful difference and testing it by using conventional statistical procedures. Spriet and Beiler (8), building on the work of Feinstein (29), demonstrated that for a given clinical trial situation in which a difference between two entities is observed, several possible conclusions are available: A substantial difference in favor of entity “A” can be demonstrated statistically, a substantial difference in favor of entity “B” can be demonstrated, or a negligible difference between them can be demonstrated (that is, equivalence). Each of these conclusions can be supported with a similar statistical strategy, namely, formulating a hypothesis to be rejected with a chosen degree of certainty (the α error or *P* value). The conclusion reached depends on the results obtained and the statistical hypothesis formulated. These authors also point out that some results may be neither large enough to be “big” nor small enough to be “small.” In such cases, no firm (statistical) conclusions about the results may be possible.

Makuch and Simon (30) demonstrated an approach of defining a maximum difference that can be accepted as “equivalent” in terms of the confidence interval around the observed difference, rather than a direct comparison of the numerical results in two groups. They derived straightforward equations for calculating the required sample size to demonstrate equivalence and applied their approach to the question of whether the “cure rate” of conservative therapy for a hypothetical case of cancer would be equivalent to that of more intensive treatment. This approach was adapted by Schuirmann (31), and a similar version was adopted by the U.S.

Food and Drug Administration for testing pharmacologic bioequivalence in the evaluation of generic drug applications (32). These techniques have been expanded upon and refined over the years (33, 34).

Among the reports in our study, about two thirds did not state a goal regarding equivalence, and equivalence was declared after a “failed” test for superiority. This approach is not satisfactory for conclusions about equivalence. The mere failure to confirm a difference statistically does not indicate the clinical importance of observed distinctions, large or small. The claim of equivalence may well be erroneous if an observed important distinction is not confirmed because the sample sizes were too small (6).

Even when the observed difference is “small,” however, absence of a quantitative boundary for a negligible or equivalent result prevents determination of the sample size needed to exclude (with reasonable certainty) the chance of a meaningful (“large”) result in a failed superiority test. The absence of a quantitative boundary also converts the decision about equivalence to an issue in statistical testing rather than scientific thought (35). This problem is illustrated by response rates of 52% (25 of 48) compared with 37% (18 of 48) in two drugs for depression (36). Despite the impressive absolute difference of 15% and the proportionate increment of 29%, the two agents were claimed to be equivalent after a conventional test of significance returned a *P* value greater than 0.05.

Even when boundaries are set for equivalence, however, inconsistencies in their choice and application can lead to contradictory results. The problem is illustrated by two trials of thrombolytic agents for myocardial infarction. First, in the International Joint Efficacy Comparison of Thrombolytics (INJECT) trial (25), two treatments were to be considered equivalent if the absolute mortality difference between investigated groups was less than 1%—a boundary chosen because a 1% mortality difference had been considered “significantly large” in the Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial (37). Second, in the Continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT) trial (38), equivalence was defined as an absolute difference in mortality of less than 0.40%. This boundary was chosen because it was the lower limit of the confidence interval around the 1% mortality difference in the GUSTO trial.

The INJECT mortality results of 9.53% compared with 9.02% led to a declaration that the two drugs were equivalent, because the absolute difference of 0.51% was less than 1.0%, the chosen equivalence boundary. In the COBALT trial, how-

ever, the mortality results were 7.98% compared with 7.54%. The absolute difference of 0.44%, although smaller than the 0.51% difference in the INJECT trial, was declared *not* equivalent because the boundary had been set at 0.40%.

Because magnitudes of “big” and “small” depend on individual judgment, the issue of whether 1.0% or 0.4% is a more appropriate boundary for an insignificant mortality difference cannot be answered definitively. However, the choice of threshold for equivalence does have important implications for study feasibility. This problem is illustrated in the COBALT trial (38), which “failed” to show equivalence. An accompanying editorial (16) pointed out that about 50 000 patients would have been required in each treatment group if an equivalence trial were designed to rule out, with 80% power, excess mortality of 0.40% for a baseline mortality rate of 7.5%.

Despite the large sample sizes sometimes required to demonstrate equivalence, equivalence trials are feasible to conduct. Twenty-two percent of the papers in our study sample successfully applied the methods we have enumerated. Even in the area of thrombolytic drug “megatrials,” rigorously designed equivalence trials have been successfully conducted. The recently reported Assessment of the Safety and Efficacy of a New Thrombolytic (ASSENT-2) study of tenecteplase compared with alteplase (39) demonstrates the application of rigorous methods in a “successful” equivalence trial. The study was designed to show the equivalence of single-bolus tenecteplase and standard alteplase (tissue plasminogen activator) treatment in acute myocardial infarction. As in the INJECT trial, the boundary for equivalence was set at a 1% absolute difference in mortality. The statistical hypothesis to be rejected was that of a 1% difference (in favor of the standard alteplase regimen) in mortality rates, with a one-sided α level of 0.05. A sample size of 16 500 was calculated on the basis of an (equal) assumed mortality rate of 7.2%. Nearly 17 000 patients were randomly assigned; the observed mortality was 6.18% for tenecteplase and 6.15% for alteplase, an absolute difference of 0.03%. The 95% confidence interval (−0.55% to 0.61%) excluded the predefined absolute difference of 1%. The *P* value of 0.006 was consistent with “equivalence” because the null hypothesis rejected by the analysis assumed superiority, whereas the alternative hypothesis assumed equivalence.

Our study has several limitations. First, because of the large number of citations (>14 000) containing the word “equivalent,” our study sample is not exhaustive. It does, however, help to demonstrate the “state of the literature.” All of the evaluative criteria were developed a priori and were applied

uniformly. Although bias due to skewing of the sample obtained is theoretically possible, it is likely that our sample overestimates methodologic appropriateness. For example, our search would miss many studies that set out to show a (large) difference, did not achieve that goal, and were reported as “failed” trials, which may be interpreted by readers as showing equivalence without a specific claim by the authors. In addition, studies in which the investigators set out to demonstrate superiority, fail to do so, and then report “equivalence” (claiming to have had that goal a priori) would be counted as having fulfilled one of our criteria; however, this would not bias our results in terms of other aspects of methodologic rigor. Moreover, methodologic standards for studies of equivalence may have improved in recent years because of increased attention to study design, but a sampling of citations from 1998 (using the same criteria as in the original study) revealed papers (for example, references 40 and 41) with the same methodologic problems.

At present, accepted stochastic boundaries (such as $\alpha = 0.05$) exist for statistical probabilities, but no descriptive boundaries have been set for “large” and “small.” These descriptive boundaries will vary with different clinical circumstances but can be chosen for each circumstance and expressed explicitly (35). Suitable procedures for setting boundaries, conducting this type of research, and reporting results can be developed as a challenge for future research.

In conclusion, the cited methodologic deficiencies may lead to harm to patients if clinically inferior treatments are erroneously deemed equivalent to a standard approach, or if potentially superior therapies are discarded as merely “equivalent.” With the increasing development and use of generic drugs, and the pressure to control medical costs by substituting less expensive therapies (or to deny therapies regarded as not cost effective), the claim that one drug, intervention, or therapy is “equivalent” to another requires close scrutiny. Attention to proper methods for conducting studies of equivalence will help avoid false claims, inconsistencies, and the inappropriate use of suboptimal therapies.

From Yale University School of Medicine, New Haven, Connecticut; and West Haven Veterans Affairs Medical Center, Veterans Affairs Connecticut Healthcare System, West Haven, Connecticut.

Grant Support: Dr. Greene was a Robert Wood Johnson Clinical Scholar at Yale University School of Medicine at the time this work was done. Dr. Concato is supported by a Career Development Award from the Veterans Affairs Health Services Research and Development Service.

Requests for Single Reprints: William L. Greene, MD, Medical Affairs Department, Genentech, Inc., Mailstop 84, 1 DNA Way, South San Francisco, CA 94080; e-mail, wgreene@gene.com.

Requests To Purchase Bulk Reprints (minimum, 100 copies): Barbara Hudson, Reprints Coordinator; phone, 215-351-2657; e-mail, bhudson@mail.acponline.org.

Current Author Addresses: Dr. Greene: Medical Affairs Department, Genentech, Inc., Mailstop 84, 1 DNA Way, South San Francisco, CA 94080.

Drs. Concato and Feinstein: Robert Wood Johnson Clinical Scholars Program, Yale University School of Medicine, 333 Cedar Street, SHM IE-61, New Haven, CT 06510.

References

- Dong BJ, Hauck WW, Gambertoglio JG, Gee L, White JR, Bubp JL, et al. Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism. *JAMA*. 1997;277:1205-13.
- Oles KS, Penry JK, Smith LD, Anderson RL, Dean JC, Riela AR. Therapeutic bioequivalency study of brand name versus generic carbamazepine. *Neurology*. 1992;42:1147-53.
- Oberlin O, Leverger G, Pacquement H, Raquin MA, Chompret A, Habrand JL, et al. Low-dose radiation therapy and reduced chemotherapy in childhood Hodgkin's disease: the experience of the French Society of Pediatric Oncology. *J Clin Oncol*. 1992;10:1602-8.
- Lefebvre JL, Chevalier D, Luboinski B, Kirkpatrick A, Collette L, Sahmoud T. Larynx preservation in pyriform sinus cancer: preliminary results of a European Organization for Research and Treatment of Cancer phase III trial. EORTC Head and Neck Cancer Cooperative Group. *J Natl Cancer Inst*. 1996; 88:890-9.
- Taggart SC, Custovic A, Richards DH, Woodcock A. GR106642X: a new, non-ozone depleting propellant for inhalers. *BMJ*. 1995;310:1639-40.
- Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
- Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-9.
- Spriet A, Beiler D. When can 'non significantly different' treatments be considered as 'equivalent'? *Br J Clin Pharmacol*. 1979;7:623-4.
- Kron IL, Kern JA, Theodore P, Flanagan TL, Haines DE, Barber MJ, et al. Does a posterior aneurysm increase the risk of endocardial resection? *Ann Thorac Surg*. 1992;54:617-20.
- Hatala R, Dinh T, Cook DJ. Once-daily aminoglycoside dosing in immunocompetent adults: a meta-analysis. *Ann Intern Med*. 1996;124:717-25.
- Stock AJ, Koford L. Therapeutic interchange of fluoxetine and sertraline: experience in the clinical setting. *Am J Hosp Pharm*. 1994;51:2279-81.
- Edgar PP, Schwartz RD. Functionally relevant gamma-aminobutyric acid A receptors: equivalence between receptor affinity (Kd) and potency (EC50)? *Mol Pharmacol*. 1992;41:1124-9.
- Palazzini M, Cristofori M, Babbini M. Bioavailability of a new controlled-release oral naproxen formulation given with and without food. *Int J Clin Pharmacol Res*. 1992;12:179-84.
- Mason KA, Thames HD, Ochran TG, Ruifrok AC, Janjan N. Comparison of continuous and pulsed low dose rate brachytherapy: biological equivalence in vivo. *Int J Radiat Oncol Biol Phys*. 1994;28:667-71.
- Cohen J. *Statistical Power Analysis in the Behavioral Sciences*. 2d ed. Hillsdale, NJ: Erlbaum; 1988.
- Ware JH, Antman EM. Equivalence trials [Editorial]. *N Engl J Med*. 1997; 337:1159-61.
- Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care*. 1989;27:S178-S189.
- Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med*. 1978;299:690-4.
- National Library of Medicine. *Abridged Index Medicus*. v 28. Washington, DC: U.S. Department of Health and Human Services; 1997.
- Rammelloo RH, Luursemma PB, Sips AP, Beumer HM, Wald FD, Cornelissen PJ. Therapeutic equivalence of a fenoterol/ipratropium bromide combination (Berodual) inhaled as a dry powder and by metered dose inhaler in chronic obstructive airway disease. *Respiration*. 1992;59:322-6.
- Jurgens RW, Downey LJ, Abernethy WD, Cutler NR, Conrad J. A comparison of circulating hormone levels in postmenopausal women receiving hormone replacement therapy. *Am J Obstet Gynecol*. 1992;167:459-60.
- Barkow D, Schwigon CD. Cefepime versus cefotaxime in the treatment of lower respiratory tract infections. *J Antimicrob Chemother*. 1993;32:187-93.
- Di Munno O, Imbimbo B, Mazzantini M, Milani S, Occhipinti G, Pasero G. Deflazacort versus methylprednisolone in polymyalgia rheumatica: clinical equivalence and relative antiinflammatory potency of different treatment regimens. *J Rheumatol*. 1995;22:1492-8.
- Harenberg J, Roebruck P, Heene DL. Subcutaneous low-molecular-weight heparin versus standard heparin and the prevention of thromboembolism in medical inpatients. *Haemostasis*. 1996;26:127-39.
- Randomised, double-blind comparison of reteplase double-bolus administration with streptokinase in acute myocardial infarction: trial to investigate equivalence. International Joint Efficacy Comparison of Thrombolytics. *Lancet*. 1995;346:329-36.
- Boissel JP, Collet JP, Lion L, Ducruet T, Moleur P, Luciani J, et al. A randomized comparison of the effect of four antihypertensive monotherapies

- on the subjective quality of life in previously untreated asymptomatic patients: field trial in general practice. The OCAP Study Group. *Optimiser le Choix d'un Anti-hypertenseur de Première Intention*. *J Hypertens*. 1995;13:1059-67.
27. Groothuis JR, Simoes EA, Lehr MV, Kramer AA, Hemming VG, Rodriguez WJ, et al. Safety and bioequivalency of three formulations of respiratory syncytial virus-enriched immunoglobulin. *Antimicrob Agents Chemother*. 1995;39:668-71.
 28. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*. 1977;33:593-602.
 29. Feinstein AR. Clinical biostatistics. XXXIV. The other side of "statistical significance": alpha, beta, delta, and the calculation of sample size. *Clin Pharmacol Ther*. 1975;18:491-505.
 30. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat Rep*. 1978;62:1037-40.
 31. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm*. 1987;15:657-80.
 32. Fuller RW, Hallett C, Dahl R. Assessing equivalence of inhaled drugs. *Respir Med*. 1995;89:525-7.
 33. Lecoutre B, Derzko G, Grouin JM. Bayesian predictive approach for inference about proportions. *Stat Med*. 1995;14:1057-63.
 34. O'Quigley J, Baudoin C. General approaches to the problem of bioequivalence. *The Statistician*. 1988;37:51-8.
 35. Feinstein AR. Zeta and delta: critical descriptive boundaries in statistical analysis. *J Clin Epidemiol*. 1998;51:527-30.
 36. Kragh-Sorensen P, Muller B, Andersen JV, Buch D, Stage KB. Moclobemide versus clomipramine in depressed patients in general practice. A randomized, double-blind, parallel, multicenter study. *J Clin Psychopharmacol*. 1995;15:245-305.
 37. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. The GUSTO investigators. *N Engl J Med*. 1993;329:673-82.
 38. A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. The Continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT) Investigators. *N Engl J Med*. 1997;337:1124-30.
 39. Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. Assessment of the Safety and Efficacy of a New Thrombolytic Investigators. *Lancet*. 1999;354:716-22.
 40. Verster GC, Joubert G, Stevens M, van der Merwe H. Generic substitution—comparing the clinical efficacy of a generic substitute for fluphenazine decanoate with the original product. *S Afr Med J*. 1998;88:260-2.
 41. Babul N, Provencher L, Laberge F, Harsanyi Z, Moulin D. Comparative efficacy and safety of controlled-release morphine suppositories and tablets in cancer pain. *J Clin Pharmacol*. 1998;38:74-81.

Personae

In an effort to bring people to the pages of *Annals*, the editors invite readers to submit photographs of people for publication. We are looking for photographs that catch people in the context of their lives and that capture personality. *Annals* will publish photographs in black and white, and black-and-white submissions are preferred. We will also accept color submissions, but the decision to publish a photograph will be made after the image is converted to black and white. Slides or prints are acceptable. Print sizes should be standard (3" × 5", 4" × 6", 5" × 7", 8" × 10"). Photographers should send two copies of each photograph. We cannot return photographs, regardless of publication. We must receive written permission to publish the photograph from the subject (or subjects) of the photograph or the subject's guardian if he or she is a child. A cover letter assuring no prior publication of the photograph and providing permission from the photographer for *Annals* to publish the image must accompany all submissions. The letter must also contain the photographer's name, academic degrees, institutional affiliation, mailing address, and telephone and fax numbers.

We look forward to receiving your photographs.

Christine Laine, MD, MPH
Deputy Editor