SPECIAL ARTICLE

Anesthesiology
2000; 92:1814-20
© 2000 American Society of Anesthesiologists, Inc.
Lippincott Williams & Wilkins, Inc.

Current Issues in Clinical Trial Design

Superiority versus *Equivalency Studies*

*Laurence Landow, M.D.**

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

HISTORICALLY, the gold standard for drug approval by the US Food and Drug Administration (FDA) has been convincing evidence of efficacy in double-blind, placebo-controlled, clinical trials. Because a placebo-controlled superiority trial provides the most straightforward opportunity for demonstrating efficacy, it is the most widely used regulatory benchmark in the drug approval process.

In some settings, a study to determine whether a drug is more efficacious than placebo may be inappropriate. The clearest example is a case in which withholding treatment or administering placebo would cause serious or irreversible harm to subjects enrolled in a clinical trial. Although no Investigational Review Board in the United States today would sanction a placebo-controlled superiority trial in men with syphilis when effective treatment is available (as occurred in the infamous Tuskegee Institute study), the National Institutes of Health recently funded studies that exposed human subjects to serious injury. In Africa and Asia, pregnant women who tested positive for the human immunodeficiency virus were randomized to the placebo group at a time when it was known that azidothymidine (AZT) prevented fetal transmission of the virus. On February 18, 1998, the placebo arm of these trials was suspended after Public Citizen[1] and members of the medical and public health communities denounced the trials as unethical.

One alternative to a placebo-controlled superiority trial is an equivalency trial.† Here, the focus is a comparison of the test drug with standard therapy (active control), not efficacy of the test drug *per se*. The primary outcome variable may be an effectiveness end point or a safety end point, *e.g.*, an adverse event, clinical laboratory variable, electrocardiographic measure, or pharmacodynamic variable.[2]

Ethical considerations aside, selecting an appropriate study design is largely dependent on the trial's objective. Because their role is to prevent ineffective or potentially harmful products from entering the marketplace, regulators primarily want to know whether an investigational drug is effective. Hence, the majority of protocols submitted to the FDA by the pharmaceutical industry are placebo-controlled superiority trials. On the other hand, clinicians want to know not only whether a new drug is effective, but how much more effective it is for their patients than current treatment options. Investigator-initiated protocols, therefore, are almost always equivalency trials that compare newly approved products with standard therapy, either for an approved indication or for an "off-label" use (unapproved indications, populations, doses, or routes of administration).[3] Drug manufacturers also conduct equivalency trials when regulatory approval is wanted for a new marketing claim; *e.g.*, intranasal administration of hydromorphone.

Even though superiority and equivalency trials share a number of features, such as blinding and randomization to minimize bias, their designs are fundamentally different. In this brief overview, I present clinical trial design issues being discussed within the regulatory community

* Instructor of Anesthesiology, Harvard Medical School, Boston Massachusetts; Former position: Medical Officer and acting Team Leader, Anesthetic and Critical Care Drugs, Food and Drug Administration, Rockville, Maryland.

Received from the Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Boston, Massachusetts. Submitted for publication November 2, 1999. Accepted for publication February 7, 2000. Support was provided solely from institutional and/or departmental sources.

Address reprint requests to Dr. Landow: 1600 Massachusetts Avenue, # 304, Cambridge, Massachusetts 02138. Address electronic mail to: landow@mediaone.net

Key words: Active control; clinical trial; equivalency trial; FDA; non-inferiority trial; placebo; superiority trial.

† Therapeutic equivalency trials as discussed here should be differentiated from bioequivalency trials, which compare bioavailability between two formulations of the same drug product, or a generic product and the original drug, to determine whether they are interchangeable. Demonstration of bioequivalency is usually regarded as tantamount to therapeutic effectiveness.

Table 1. Usefulness of Specific Control Types in Various Situations*

Trial Objective	Type of Control/Design			
	Placebo-control Superiority Trial	Active-control Superiority Trial	Active-control Noninferiority Trial	Three-arm Placebo + Active-control Noninferiority Trial
Measure absolute effect size	Yes	No	No	Yes
Show existence of effect	Yes	Yes	Possibly	Yes
Show dose–response relationship	No	No	No	No
Compare therapy	No	Yes	Possibly	Yes
Show assay sensitivity	Yes	Yes	No	Yes

* Modified from Draft Guidelines on Statistical Principles for Clinical Trials: Notice of Availability.[2]

and cite examples that are relevant to anesthesiology and critical care medicine.

Hypothesis Testing in Superiority Trials

Superiority trials are designed to show a treatment difference or "effect" between a test drug and a control (table 1). The control may be either placebo (the so-called "classic" superiority trial) or active control (standard of care). In a superiority trial comparing a test drug with an active control, the difference between the two drugs is always smaller, often much smaller, than the expected difference between drug and placebo, resulting in the need for larger sample sizes.[2]

The format of a superiority trial can be expressed by two hypotheses: the null hypothesis (H_0), which states that there is no difference between the test drug and control in terms of some outcome variable, and the alternate hypothesis (H_A), which states that there is a difference. For the purposes of regulatory approval, effectiveness is shown when the difference between the observed treatment effect of the test drug compared with that of the control exceeds some prespecified threshold considered to be "clinically relevant."

In 1998, Glaxo-Welcome (Triangle Park, NC) submitted a protocol to the FDA for a double-blind, placebo-controlled, phase III multicenter superiority trial to test whether administration of L-N^G-methylarginine hydrochloride (546C88) resulted in a statistically significant reduction in 28-day mortality in patients with septic shock. In this trial, the null and alternate hypotheses were defined as follows:

H_0: 546C88 does not reduce the 28-day mortality rate
H_A: 546C88 reduces the 28-day mortality rate.

In addition to an unambiguous primary outcome variable (28-day mortality rate), this protocol contained a number of features found in well-designed clinical trials: (1) a clearly stated objective, (2) strict inclusion and exclusion criteria; (3) blinding and randomization techniques; (4) composition of the Data Safety Monitoring Board, timing of an interim data analysis, and criteria for stopping the trial prematurely; (5) a power analysis, *i.e.*, an estimate of the sample size necessary based on published survival rates in patients with septic shock; (6) the type I error rate (likelihood of finding a reduction in mortality that could have been a result of chance—typically, 0.05 or less) and the type II error rate (likelihood of not finding a treatment effect when one actually exists—typically 0.20 or less); and (7) the statistical model for analyzing "drop-outs" (subject withdrawals), covariates (age, gender, physiologic status), and protocol violations. To determine whether 546C88 was effective, the sponsor proposed (and the agency agreed) that confidence intervals for the two groups be constructed and a "win" declared if there was a reduction of greater than 10% in the 28-day mortality rate, based on a statistically accepted measure (likelihood ratio test). Regrettably, the trial had to be discontinued early because an interim safety analysis revealed an unacceptable increase in mortality in the 546C88 group.

Some of the early exploratory studies designed to assess the relative potency of intravenous morphine and oral transmucosal fentanyl citrate (OTFC; Actiq; Anesta, Salt Lake City, UT) for "breakthrough" cancer pain also involved double-blind, placebo-controlled superiority designs. Opioid-naive postoperative surgery patients with access to patient-controlled intravenous morphine (rescue medication) were randomized to receive placebo or OTFC on a fixed dosing schedule. Not surprisingly, the placebo group required significantly more rescue medication (the study endpoint) than the test drug group, thereby showing the efficacy of the test drug.

Figure 1 depicts the results of three hypothetical superiority trials (A, B, C) in which three different drugs (A, B, C) are compared with a placebo for treatment of

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

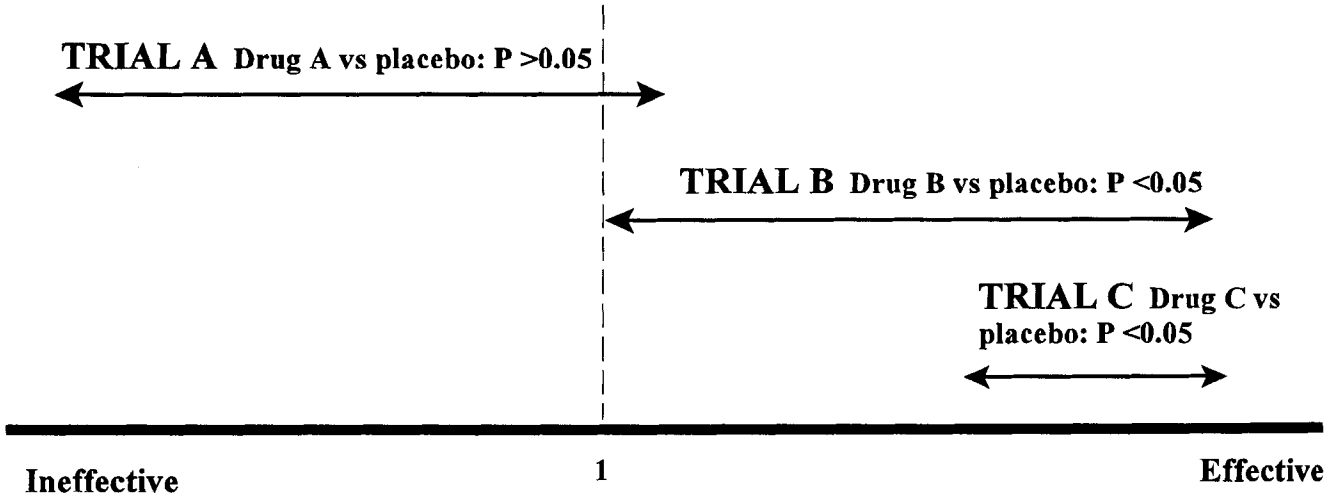

Fig. 1. 95% Confidence intervals for three, hypothetical, placebo-controlled, superiority trials. When the confidence interval crosses 1, the intervention is no different from placebo (drug A). When the confidence interval does not cross 1 and $P < 0.05$, the intervention is effective (drugs B and C). When an intervention is effective but the 95% confidence interval is wide (drug B), there is less reassurance than when it is narrow (drug C) that the same treatment effect will be observed in the total population as in the study population.

the same disease. As the figure shows, one method of summarizing data is through the use of *P* values: the smaller the *P* value, the more likely it is that the null hypothesis is false. Another, more informative approach to assessing the credibility of a clinical outcome is the size of the confidence interval—narrow intervals (little physiologic variability or "noise") providing more reassurance than wide ones that a comparable difference in treatment effect will be observed in the general population once the drug is marketed.[4] It is important to note that even if the treatment effect is constant across two or more studies ("treatment homogeneity"), this does not necessarily imply that treatment homogeneity will be observed subsequently.[5] Some analgesics and antidepressants are notorious for showing an effect in early trials but failing to show this effect in subsequent studies. Explanations to account for this "treatment heterogeneity" include variance in response rates within subpopulations, selection of different endpoints or different time points, and unrecognized subject selection bias.

Hypothesis Testing in Equivalence–Noninferiority Trials

In trials designed to test equivalence (or, as is more often the case, noninferiority‡), one seeks to reject the alternate hypothesis that there is a difference between two products, *i.e.*, discover how much worse drug B can be than drug A and still be acceptable (table 1). This can be a difference in efficacy or safety; for example, atracurium and cisatracurium are both effective muscle relaxants, yet the latter may be advantageous in clinical settings in which histamine release is undesirable.

The magnitude of this clinically acceptable difference (designated by the Greek letter δ) must be justified in the protocol and accepted by the FDA review team before the trial gets underway. In practice, determination of δ is a function of several factors: results of previous studies in the same population, clinical importance of the claimed benefits of the test drug, and the clinical judgment of the medical reviewer. In some cases, a clinically acceptable difference may be smaller than the "clinically relevant" difference found in superiority trials designed to show that a difference exists.

In one marketing application submitted to the FDA, the sponsor wanted to demonstrate in patients undergoing open heart surgery in association with cardiopulmonary bypass that Bretschneider cardioplegia solution

‡ In an active control equivalency trial, both the upper and the lower equivalence margins are needed; equivalence is inferred when the entire confidence interval falls within the equivalence margins. In a noninferiority trial, the finding of interest is one sided, so only the lower boundary is needed.

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

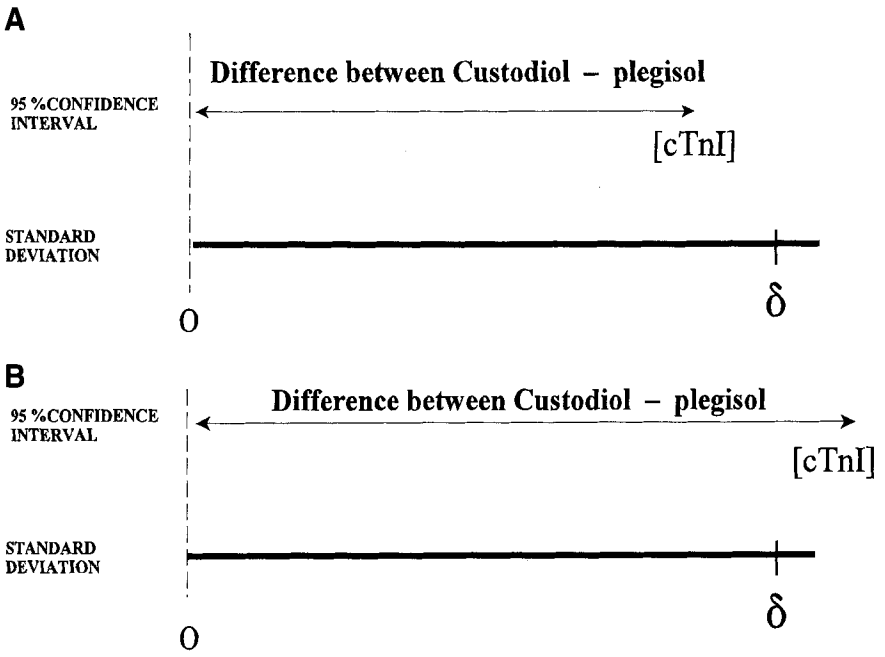

Fig. 2. 95% Confidence interval (thin solid line) and SD (thick solid line) for a hypothetical noninferiority trial comparing Custodiol and plegisol (active control), using serum area under the curve troponin I concentration ([cTnI]) as the primary outcome variable (treatment effect). The confidence interval is calculated from the difference between the treatment effect of the two cardioplegia solutions. (*A*) Shows noninferiority of the two solutions; *i.e.*, the upper end of the confidence interval is 0.5 SD or less (δ). (*B*) Depicts the converse (see text).

(Custodiol; Köhler Chemie GmbH, Alsbach-Hähnlein, Germany) was as effective as plegisol, the only FDA-approved cardioplegia solution. A surrogate of myocardial protection, serum troponin I concentration ([cTn]), was proposed as the primary efficacy variable. The agency indicated that the new solution would be approved if clinical trials showed noninferiority, *i.e.*, showed that the confidence interval of the difference in the area under the curve [cTnI] was no more than 0.5 SD (δ) higher in subjects treated with Custodiol than in those treated with plegisol (fig. 2).

In another submission, the sponsor (Organon, West Orange, NJ) wanted to show that the percentage of subjects demonstrating clinically acceptable intubating conditions (rated "good to excellent" using the Viby-Mogensen scoring system) 60 s after intravenous administration of the nondepolarizing muscle relaxant Org 9487 (rapacuronium, Raplon; Organon) was equivalent to that among subjects receiving succinylcholine. In statistical shorthand, the alternate and null hypotheses were expressed as

H_A: percentage of Org 9487 patients − percentage of succinylcholine patients $> \delta$

H_0: percentage of Org 9487 patients − percentage of succinylcholine patients $\leq \delta$

where δ was prespecified as 10%. The agency indicated that Organon would be allowed to make this marketing claim if the clinical trials showed that the upper bound on the inferiority end of a 95% confidence interval for the between-group difference was small enough to be clinically insignificant (here, ≤ 10%).

These examples underscore a number of points. First, the protocol should clarify ahead of time whether one- or two-sided tests of statistical significance will be used and, in particular, justify prospectively the use of one-sided tests. Second, the active control and its dosage should be selected with care. A suitable choice is an agent in widespread use for which efficacy against placebo for the relevant indication has been clearly established and quantified in well-designed and well-documented superiority trials, and one that would be expected to exhibit similar efficacy reliably (in terms of some prespecified magnitude) in the contemplated active control study, had placebo been present. Third, and most important, in noninferiority trials in which one compares an investigational drug with an active control, failure to find a difference does not necessarily mean there is no difference, as will be discussed in the next section.

Problems Encountered in Equivalency–Noninferiority Trials

Assay Sensitivity

Assay sensitivity refers to the ability of a specific trial to detect differences between treatments, if they exist. The

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

FDA Director of the Office of Medical Policy Robert Temple has stated, "If we cannot be very certain that the positive (active) control in a study would have beaten a placebo group, had one been present, the fundamental assumption of the positive control study cannot be made and that design must be considered inappropriate."[6]

The active controls selected for the Custodiol and Org 9487 clinical trials (plegisol and succinylcholine, respectively) clearly satisfy Temple's criterion. In clinical settings in which no gold standard treatment exists and in which event rates can vary widely, trial designs without placebo control are unlikely to convincingly show effectiveness.

In a recent meta-analysis of 33 randomized, controlled clinical trials, comprising 4,872 subjects, that studied the antiemetic effectiveness of ondansetron,[7] there were eight different regimens with 28 different comparators, including metoclopramide (6 trials), droperidol (11 trials), and metoclopramide + droperidol (1 trial). Of note, only 19 of the trials included a placebo arm; in these, nausea or vomiting rates in the placebo group varied between 1 and 80% for outcomes up to 6 h after surgery and between 10 and 96% for outcomes up to 48 h after. Many of the trials showed no difference between ondansetron and active control.

The only conclusions that can be reached when two drugs show a similar treatment effect are (1) both drugs are effective to a similar degree; (2) both drugs are equally ineffective; or (3) the trial is underpowered; *i.e.*, in the face of a defined event rate, the sample size is too small to show that a real difference exists between two treatments. In fact, the only time one *can* be sure that a noninferiority trial can differentiate a real difference is when it rejects the claim of noninferiority. (According to Temple, "There is no such thing as equivalence in [clinical] trial design. All one can ever say is the difference is greater than thus-and-such.")[8]

To draw correct conclusions in noninferiority trials, the test drug and active control both must be shown to be effective in the same population, for the same endpoint, and at roughly the same time point; the only way to ascertain this is with a trial that can detect a difference between drug and placebo, if it exists, by concurrently measuring the placebo response. As alluded to previously (treatment heterogeneity), there is an often an unstated—but not always recognized—assumption that the active drug is effective in the particular study in question, which is not necessarily true.[9]

Temple has highlighted an additional problem with noninferiority trials.[10] In trials intended to show superiority, there is a strong imperative to minimize "sloppiness" in the design and conduct (*e.g.*, weak enforcement of inclusion-exclusion criteria, lack of adequate follow-up, excessive variability of measurements, inadequate blinding) because it increases the likelihood of failing to show a difference between treatments when one exists. The stimulus to engage in these efforts in a noninferiority trial is much weaker because sloppiness tends to "dilute" or reduce observed differences between groups.[2] For example, the sponsor of a new drug might select a subgroup of patients in whom, or a time point or dosage at which, the treatment effect in previous trials with active control was small, thereby making it easier to show equivalence. Readers interested in an opposing view of this topic should review the article by Hauck and Anderson.[11]

Trial Designs to Protect Human Subjects from Harm

As implied in the preceding section, the agency views noninferiority trials as potentially problematic because they do not measure efficacy directly. One solution to this problem is the addition of a third placebo arm (table 1). To some observers, adding a placebo arm in, say, an antiemetic drug trial is unethical. The problem with this argument is that exposing human subjects to a product of unproven benefit and uncertain safety, and in a trial destined to produce unreliable results, is itself unethical.[12] Conversely, when there is serious concern that inclusion of a placebo arm will be life-threatening, result in irreversible morbidity, or cause gratuitous pain and suffering, consideration should be given to the following design modifications.[13]

(1) Historical control trials, in which differences in treatment effect between test drug and historical control are used as a basis for regulatory approval. Here, it is critical to carefully review the design and conduct of previous studies on a trial-by-trial basis in terms of inclusion and exclusion criteria, dosage and regimen of therapy, outcome measures, and follow-up. The obvious weakness in this design is that the historical event rate may have evolved substantially over time because of breakthroughs in standard of care and diagnosis and broad changes in diet.

(2) Add-on studies, in which both treatment groups continue to receive standard treatment so that therapy is not withheld from a population known to benefit from it. Then, one group is randomized to receive the test drug (which must be of a different

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

pharmacologic class than standard treatment) and the other the placebo:

treatment A *versus* treatment A + treatment B

For instance, Fujii *et al.*[14] found that granisetron (which "beat placebo" in previous trials) + saline was less effective than granisetron + dexamethasone in preventing postoperative emesis in children undergoing strabismus repair or tonsillectomy with or without adenoidectomy.

Conceptually, the strategy underlying an add-on study is that the size of the difference in effect between an effective drug (B) and no treatment is likely to be greater than between two effective drugs (A + B). This argument assumes, of course, that drug B can provide additional benefit, *i.e.*, a "ceiling effect" has not already been reached using drug A alone.

(3) "Enrichment" studies: Enrichment refers to enrolling only those subjects who demonstrate a favorable—or, in cases in which safety is an issue, unfavorable—response to an investigational drug, thereby producing a population more likely to discriminate between an active and an inactive therapy.

An add-on enrichment trial design was used in one of the OTFC trials for breakthrough cancer pain, which followed previous trials designed to determine the best way to define the successful dose of OTFC. This was a multicenter, double-blind, placebo-controlled, crossover study of subjects prescribed stable around-the-clock opioid therapy for chronic cancer pain, who also required additional analgesia for episodes of breakthrough pain. In the open-label phase of the trial, subjects identified an effective dose of OTFC by titration through the available dosage strengths (200–1600 μg). Those who were titrated to a single-dosage strength that provided adequate pain relief with acceptable side effects for breakthrough episodes ("responders") entered the double blind phase in which they each received 10 prenumbered OTFC units, of which 7 were their effective dose and 3 were placebo. Subjects were asked to record pain intensity, pain relief, global performance of the treatment, and adverse events.

(4) Randomized "withdrawal trials" in which subjects in the investigational drug arm of a clinical trial are randomly assigned to continued treatment with the investigational drug or to placebo. Any difference that emerges between groups shows the effect of the active treatment, even though there is no direct assessment of the absolute treatment effect. The advantage of such a design, when used with an early-escape end point (such as return of symptoms), is the short duration of exposure to placebo. One setting in which withdrawal trials are attractive is in patients with angina, in whom long-term randomization to placebo would be unethical. Randomized withdrawal designs can also assign subjects to multiple dosage levels of test drug to determine the most effective dose.

(5) Replacement trials, in which the test drug (at several different doses) or placebo is added by random assignment to standard treatment administered at an effective dose, followed by tapered withdrawal of the conventional treatment. The ability to maintain the subjects' baseline status is then observed in the drug and placebo groups using prespecified success criteria. This approach has been used to study steroid-sparing substitutions in steroid-dependent patients without the need for initial steroid withdrawal and recrudescence of symptoms in a "wash-out" period.

(6) "Putative placebo" trials: A putative placebo is the current standard of care (*e.g.*, aspirin administration in postmyocardial infarction patients), the effect of which is of such magnitude, consistency, and demonstrated benefit (effectiveness) when compared with placebo, that it is unethical to withhold it from a subject in a clinical trial. To be successful, the test drug must show an effect that is superior, not necessarily to active control (the putative placebo), but to the best outcome that might have been seen with placebo if placebo had been present.[15] Obviously, putative placebo trials are not appropriate in situations in which the test drug does not consistently beat placebo.

Conclusion

Innovative technologies are revolutionizing the drug discovery process, resulting in an exponential increase each year in the number of new drugs that enter the pharmaceutical industry's pipelines. Already, regulators are coming under pressure to accept more noninferiority trials because of the plethora of effective products available and appeals from clinicians and their patients for studies that reflect clinical practice.

Unless ethically prohibited, drug manufacturers and clinical investigators should be strongly encouraged to include a third placebo arm in their noninferiority effi-

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024

cacy trials so that their results will answer the questions of all parties concerned.

The author thanks Bill Camann, M.D., Anesthesiology Department, Brigham and Women's Hospital, Boston, Massachusetts, for his thoughtful comments.

References

1. Lurie P, Wolfe SM: Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. New Engl J Med 1997; 337:853-6

2. Draft Guidelines on Statistical Principles for Clinical Trials: Notice of Availability. International Conference on Harmonization. Federal Register 1997; 62:25711-26

3. Landow L: Off-label use of approved drugs. Chest 1999; 116: 589-91

4. Longford NT, Nelder JA: Statistics versus statistical science in the regulatory process. Stat Med 1999; 18:2311-20

5. Longford NT: Selection bias and treatment heterogeneity in clinical trials. Stat Med 1999; 18:1467-74

6. Temple RJ: Difficulties in evaluating positive control trials. Proceedings of the biopharmaceutical section of the American Statistical Association. 1983:1-7

7. Tramer MR, Reynolds DJM, Moore RA, McQuay HJ: When placebo controlled trials are essential and equivalence trials are inadequate. BMJ 1998; 317:875-80

8. Cardiovascular and Renal Drugs Advisory Committee Transcript [online]. October 23, 1997:54; Available at http://www.FDA.gov/cder/foi/special/99/case-trans-42199.txt

9. Committee for Advanced Scientific Education Seminar Series. The Use of Placebos in Clinical Trials and the Ethics of the Use of Placebos. April 21, 1999; Available at http://www.FDA.gov/cder/foi/special/99/case-trans-42199.txt

10. Temple RJ: Government viewpoint of clinical trials. Drug Information J 1982; 16:10-7

11. Hauck WW, Anderson S: Some issues in the design and analysis of equivalence trials. Drug Information J 1999; 33:109-18

12. Collier J: Confusion over use of placebos in clinical trials. BMJ 1995; 311:821-2

13. Temple RJ: Special study designs: Early escape, enrichment, studies in non-responders. Communications Statistics 1994; 23:499-531

14. Fujii Y, Tanaka H, Toyooka H: Granisetron and dexamethasone provide more improved prevention of postoperative emesis than granisetron alone in children. Can J Anaesth 1996; 43:1229-32

15. Borer JS: Future direction of cardiovascular drug development, Cardiovascular Drug Development: Protocol Design and Methodology. Edited by Borer JS, Somberg JC. New York, Marcel Dekker, 1999, p 216

Downloaded from http://asa2.silverchair.com/anesthesiology/article-pdf/92/6/1814/429786/0000542-200006000-00042.pdf by guest on 09 April 2024