Interpreting Results of Clinical Trials: A Conceptual Framework

Ajay K. Singh,* Ken Kelley,† and Rajiv Agarwal‡

*Renal Division, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; and †Inquiry Methodology Program and ‡Division of Nephrology, Indiana University, Indianapolis, Indiana

Clin J Am Soc Nephrol 3: 1246–1252, 2008. doi: 10.2215/CJN.03580807

Clinical trials are generally designed to test the superiority of an intervention (*e.g.*, treatment, procedure, or device) as compared with a control. Trials that claim superiority of an intervention most often try to reject the null hypothesis, which generally states that the effect of an intervention of interest is no different from the control. In this editorial, we introduce a conceptual framework for readers, reviewers, and those involved in guideline development. This paradigm is based on evaluating a study on its statistical merits (result-based merit) as well as the clinical relevance of the potential treatment effect (process-based merit). We propose a decision matrix that incorporates these ideas in formulating the acceptability of a study for publication and/or inclusion in a guideline. Although noninferiority trials and equivalence trials are other valid trial designs, here we largely focus our discussion on superiority trials.

Studies termed "negative" are commonly defined as those where the difference for the primary endpoint has a P value greater than or equal to 0.05 ($P \geq 0.05$) (1), that is, where the null hypothesis is not rejected. These studies are difficult to publish because they are said to be "nonsignificant." In other words, the data are not strong enough to persuade rejection of the null hypothesis. A high P value is frequently interpreted as proof that the null hypothesis is true; however, such an interpretation is a logical fallacy. A nonsignificant result implies that there was not enough evidence to infer probabilistically that the null hypothesis can be rejected. What is important to keep in mind is that the absence of evidence does not imply evidence of absence (2,3). On the other hand, if a small P value is observed, it implies there is evidence that the null hypothesis is false, which is why much stronger clams can be made when the null hypothesis is rejected. Recall that the null hypothesis (H_0) is a stated value of the population parameter that is set up to be refuted. Most often, the value of H_0 states that the effect of interest (*e.g.*, mean difference, squared multiple correlation coefficient, regression coefficient) is zero. However, this need not be the case. This point is illustrated in Table 1, a 2 × 2 statistical inference decision table, wherein what is the true but unknown state of the world is crossed with the statistical decision, thereby generating conceptual definitions for the type 1 and type 2 error rates.

In "reject-support" hypothesis testing, by far the most common scenario, H_0 is not generally what the researcher actually believes (4), and thus the value of H_0 is generally set up to be refuted. When H_0 is refuted, that is when $P < \alpha$ (*e.g.* $P < 0.05$), strong support exists for rejecting the null hypothesis in favor of the alternative hypothesis, where α is the type I error rate. Until statistical evidence in the form of a hypothesis test indicates otherwise, the null hypothesis is presumed true (1). For example, in a clinical trial with an intervention and a control group, the null hypothesis generally proposes that the intervention and control are equally as effective or ineffective. That is, the population mean of the treatment and control groups is assumed to be the same until an effect estimate (which reaches some prespecified statistical threshold) is observed.

What should be clear is that a large P value does not in any way *prove* that H_0 is true. When H_0 is not rejected, it implies that either H_0 is actually true, reflecting a correct decision (lower left cell of Table 1), or that H_0 is actually false but that there was not enough evidence observed to reject H_0, a type II error (lower right cell of Table 1). When H_0 is rejected, it implies that either H_0 is actually false, reflecting a correct decision (upper right cell of Table 1), or that H_0 is actually true, but the evidence observed is "unlikely" (which will occur with probability α when the null hypothesis is true; upper left cell of Table 1).

The Importance of Confidence Intervals

When a study demonstrates a P value $\geq \alpha$ (*e.g.*, α = 05), it is important to ask two questions: "What is the $(1 - \alpha)100\%$ confidence interval (CI) for the population effect of interest?" "Was there sufficient power to detect a clinically meaningful effect if it in fact existed?" Moreover, even when the null hypothesis is rejected, CIs provide important insights.

Reporting of CIs is widely regarded by statisticians and methodologists as the optimal way to report and interpret the results of research studies because CIs convey information regarding direction, magnitude, and accuracy of effects (5–8). If the entire span of a 95% CI contains effects that are clinically or scientifically trivial (*e.g.*, mean differences, relative risks, proportion of variance accounted for), researchers can then confi-

Published online ahead of print. Publication date available at www.cjasn.org.

Correspondence: Dr. Ajay K. Singh, Renal Division, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115. Phone: 617-732-5951; Fax: 617-732–6392; E-mail: asingh@partners.org

ISSN: 1555-9041/305–1246

Table 1. Statistical inference decision table

Inferential Decision	Unknown but True State of the World	
	H_0 True	H_0 False
Reject H_0	Type I error Probability = a	Correct decision Probability = 1 − b
Do not reject H_0	Correct decision Probability = 1 − a	Type II error Probability = b

a, type I error rate; b, type II error rate (1 − b is power).

dently state at best the intervention does not provide clinically or scientifically meaningful results, even if the results happen to reach a prespecified statistical threshold. On the other hand, if the CI is wide enough to contain both the null value as well as values that could be clinically or scientifically important, then the study is inconclusive at best. Indeed, CIs provide a framework to evaluate studies on their result-based *versus* process-based merits (or nonstatistics merits) (Table 2; Figure 1). From the perspective of result-based merit, studies that fall in cells "a" and "c" arguably merit publication, whereas those in "b" and "d" do not. Of course, the most valuable study would be in category "a," one that demonstrates a significant effect and narrow CIs. Often, journals tend to accept studies in "a" and reject those in "c" and "d." Studies in category "b" tend to take on a hit-or-miss character. High impact journals have a high threshold to accept studies in "b" because a wide CI is not very informative; the real treatment effect may be large (or small) or even clinically trivial. Although studies in category "b" do have modest result-based merit, their attractiveness can be advanced if the investigators are able to convince the journal of the clinical relevance of the potential treatment effect (process-based merit). On the other hand, studies in cell "d" should not be automatically rejected because a wide CI may encompass a clinically important treatment effect. These studies are frequently the most problematic because trials with small sample sizes are likely to fall into this category nonsignificant effects and wide CIs and, if published, may provide little incentive to obtain large sample sizes at best, or misinform at worst. Some of these studies may be clinically informative, that is, they have high clinical relevance, because they generate hypotheses that could be tested in subsequently larger studies, or stimulate ancillary analyses that might influence clinical practice These studies have limited "results-based" merit but may have high "process-based merit" (Figure 1, bottom right box).

Wide CIs may detract from accurately estimating the population value. This may generally reflect inadequate sample size. When the sample size is too small, there is a high degree of variability and thus imprecise estimation. Performing power analyses and reporting information on analytic processes taken to estimate sample size have gained much credence, and indeed, are required by many funding agencies when proposing studies (9–13). As Goodman and Berlin (10) have pointed out, studies with low statistical power have sample sizes that are too small to reliably detect an effect (*i.e.*, reject H_0) if indeed such an effect exists. Freiman *et al.* analyzed the statistical power of 71 studies with $P > 0.05$ that compared two treatments (5). For most of the studies, the power to detect a 50% improvement was low. Sample size can be planned before initiating a trial so that the expected width of the CI is sufficiently narrow, increasing the likelihood of producing an estimate that closely corresponds to its population value (10,12), an approach that has been termed "accuracy in parameter estimation."

To illustrate these points, it is worth examining a recently published study on the optimal hemoglobin target in chronic kidney disease patients. In the Cardiovascular Risk Reduction by Early Anemia Treatment with Epoetin β (CREATE) study

Table 2. A results-based approach to studies

	CI, Narrow	CI, Wide
H_0, reject	a	b
H_0, fail to reject	c	d

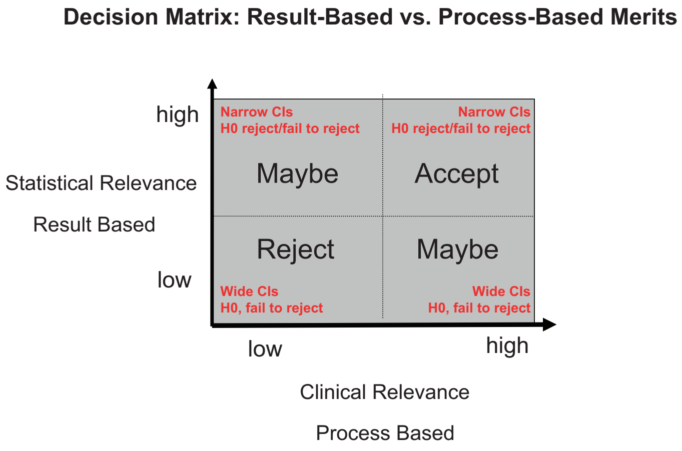

Figure 1. A decision matrix for readers, reviewers, and guideline makers to conceptualize results-based and process-based merits in clinical trials. Studies that have high results-based merit are characterized by narrow CIs and may or may not reject H0. Their level of acceptability, particularly in high impact journals, should be viewed through the lens of process-based merit. On the other hand, studies that are characterized by both low results-based and process-based merit are frequently rejected.

(14), the authors conclude "the complete correction of anemia did not affect the likelihood of a first cardiovascular event (58 events in group 1 *versus* 47 events in group 2; hazard ratio, 0.78; 95% CI 0.53 to 1.14; $P = 0.20$." In this study, the hazard ratio of 0.78 favored lower risk for subjects assigned to the lower hemoglobin group. Although this study did not yield statistically significant results, there is also no evidence that the H_0 is true (*i.e.*, what some might term a "negative" study). Indeed, technically, H0 can never be proven true; this is particularly so in the context of a superiority trial. Going beyond the point estimate alone, it is clear that the span of the CI was modestly wide, and the hazard ratio favoring increased risk, could be clinically important. The most plausible explanation of the nonsignificant *P* value is provided in the statistical section of the paper. The authors calculated that 600 patients would be needed to yield statistical power of 80% to detect a 33% reduction in the primary endpoint between the higher and lower target hemoglobin groups, given an annual incidence of the primary endpoint of 15% among patients in the lower hemoglobin group at a significance level of 5%. However, the study had a much lower rate of cardiovascular events than anticipated (5% observed *versus* 15% anticipated). Only 105 events of the expected 200 events were observed (the investigators calculated that 200 events would be needed for 80% power given their *a priori* estimates of the anticipated population characteristics). Thus, a reasonable conclusion for the CREATE study is that it is inconclusive because the 95% CI contains clinically important values, a much different conclusion than stating that there is no difference in correction of anemia on the likelihood of a first cardiovascular event. The CREATE study represents a category "c" or "d" study null findings and modestly wide CIs (modest results-based merit). However, it is highly clinically relevant and has high process based merits (Figure 1, bottom right box), justifying its publication in a high impact journal.

Another illustrative example is the largest trial in hypertension ever conducted, the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (15). A total of 33,357 participants aged 55 yr or older with hypertension and at least 1 other coronary risk factor from 623 North American centers were randomly assigned to chlorthalidone, 12.5 to 25 mg/d (n = 15 255); amlodipine, 2.5 to 10 mg/d (n = 9048); or lisinopril, 10 to 40 mg/d (n = 9054) for planned follow-up of approximately 4 to 8 yr. The primary outcome was combined fatal coronary heart disease or nonfatal myocardial infarction, analyzed by intent-to-treat, which occurred in 2956 participants. Compared with chlorthalidone (6-yr rate, 11.5%), the relative risks (RRs) were 0.98 (95% CI, 0.90 to 1.07) for amlodipine (6-yr rate, 11.3%) and 0.99 (95% CI, 0.91 to 1.08) for lisinopril (6-yr rate, 11.4%). Although there were no statistically significant effects for the RRs, the narrow CIs arguably contain plausible values for the population RR that some might view as clinically trivial. That is, the study showed support for the null hypothesis in the sense that the smallest and largest plausible values of the population RRs might be regarded as clinically trivial. (Even here, however, accepting the null hypothesis is controversial because a 10% reduction in cardiovascular events may be clinically important in this population) the narrow CI and the high clinical relevance warrant placing it in the top right hand corner of the decision matrix (Figure 1).

Knowing what transpired during a study is also important in judging whether the study result is compatible with a type II error. The conclusions drawn from the Normalization of Hematocrit Study (NHS) (16), have been questioned by some (17). In this study, 183 deaths and 19 first nonfatal myocardial infarctions were recorded among patients assigned to the higher hematocrit group *versus* 150 deaths and 14 nonfatal myocardial infarctions among those in the low hematocrit group. The relative risk was 1.3 and the 95% CI limits were 0 and 1.9. Here the issue was not power *per se*, but that the study was halted by the data safety monitoring board even though the difference in the event-free survival between the 2 groups did not reach the prespecified stopping boundary before the study was stopped. The RR of 1.3 and the span of the CIs should give pause. So should the statement by the authors in the discussion section: "Our study was halted when differences in mortality between the groups were recognized as sufficient to make it very unlikely that continuation of the study would reveal a benefit for the normal-hematocrit group and the results were nearing the statistical boundary for a higher mortality rate in the normal hematocrit group" (16). NHS, as published, was not a negative study but a statistically inconclusive study. However, the 95% CIs for the RRs (0.9, 1.9) suggest risk, and the results have potentially important clinical implications. The NHS is a category "d" study (Table 2) but an important trial that was worthy of publication in a high-impact journal on its process-based merits (bottom right hand corner, Figure 1).

Bias in clinical trials may also affect the results-based merit of a trial. Bias may threaten the internal validity of a trial and may generate erroneous conclusions. Potential sources of bias are listed in Table 3. Even well-designed trials may be vulnerable to the possibility of bias. To illustrate this point, consider the study by Ginzler *et al.* (18) that evaluated the effect of mycophenalate mofetil *versus* cyclophosphamide as induction therapy for the treatment of lupus nephritis. The study represents a category "b" study (Table 2). A total of 140 patients were recruited: 71 randomly assigned to mycophenalate mofetil and 69 to receive cyclophosphamide. In the intention-to-treat analysis, 16 of the 71 patients (22.5%) receiving mycophenolate mofetil and 4 of the 69 patients receiving cyclophosphamide (5.8%) had complete remission, an absolute treatment difference of 16.7% (95% CI, 5.6% to 27.9%; $P = 0.005$). However, scrutiny of the withdrawal and crossover rates from one group to the other raises the possibility of both withdrawal bias and follow-up bias. Notably, 7 patients (9.9%) assigned to mycophenalate mofetil and 15 patients (21.7%) assigned to cyclophosphamide were lost to follow-up. Furthermore, 24 patients were withdrawn from the study: 9 in the mycophenolate mofetil group and 15 in the cyclophosphamide group. There was greater noncompliance and a higher crossover rate in the cyclophosphamide compared with the mycophenalate mofetil group. Whether mycophenalate mofetil is actually superior to cyclophosphamide will need to be more definitively confirmed in a larger randomized trial.

Table 3. Potential sources of bias in a clinical trial

Types of Bias	Potential Effects of Bias
Intervention assignment bias, *e.g.*, unconcealed allocation or unblinding	Knowledge of assignment might lead to performance bias, *e.g.*, extra attention to participants for one of the intervention groups than the other; unblinding of participants to interim study results might result in performance bias
Exclusion bias (exclusions after randomization)	Nonrandom withdrawal in intervention groups might result in intervention groups becoming less comparable than at randomization
Performance bias	Intervention groups might receive cointerventions differentially (*e.g.*, greater counseling or follow-up for one group than the other)
Differential compliance bias	Subjects in the two arms of the trial may have different levels of compliance with the assigned therapy; this might dilute the intervention effect; or subjects who cross over from one arm to the other might dilute the intervention effect
Follow-up bias	More subjects in one arm lost to follow-up than in the other arm; or differential follow-up between the intervention and control groups?
Measurement bias	Inaccuracy of assessment method especially if nonrandomly distributed between intervention groups
Detection bias	Differential rates of outcomes assessment between the intervention and control groups
Selective reporting of results	Conflict of interest because of industry sponsorship with selective reporting of favorable results or lack of reporting of unfavorable data

Another question is whether to weigh limitations in design that likely result in a P value of >0.05. The Modification in Diet in Renal Disease (MDRD study) (19) provides an example of such an issue, and as such represents a category "d" study. MDRD comprised a 2×2 factorial design to test the utility of protein restriction or aggressive blood pressure (BP) control in patients with chronic kidney disease. Patients were stratified by glomerular filtration rates (GFRs) in what we refer to here as study 1 and study 2. Study 1 randomized 585 subjects with GFRs of 25 to 55 ml/min per 1.73 m^2 to a usual protein or a low protein diet (1.3 or 0.58 g of protein per kilogram of body weight per day) and to a usual or a low-blood-pressure group (mean arterial BP of 107 or 92 mmHg). In study 2, 255 subjects with GFRs of 13 to 24 ml/min per 1.73 m^2 were randomly assigned to a low protein (0.58 g protein/kg body weight) or a very-low-protein diet (0.28 g protein per kg body weight) with a keto acid/amino acid supplement and within each strata to a usual or low BP with BP targets as in study 1. A key limitation in MDRD was an unanticipated lower rate of decline in GFR effectively limiting the potential benefits of the intervention. The authors speculated that a longer follow-up might have detected a difference in the GFR between the treatment groups. A second limitation in MDRD was an unexpectedly steep decline in the GFR in subjects randomized to a low protein diet, immediately following exposure to the intervention compared with those prescribed the usual protein diet. According to the authors, this was not anticipated based on either prior studies or a feasibility study. The authors speculated that this might have reflected a hemodynamic response to the protein restriction. Therefore, it is possible that this factor also contributed to the nonsignificant results. Turning again to the CIs in evaluating the P value, the mean decline in the low protein group was 1.2 ml/min less and the low pressure group was 1.6 ml/min less. The 95% CI of the difference in the protein group was -1.1 and 3.6 ($P = 0.30$) and for the BP group -0.8 and 3.9 ($P = 0.18$). The wide span of the CIs, especially in the context of the unanticipated nature of the limitations, supports MDRD as inconclusive but with process based merits. MDRD provided value from a clinical standpoint because of its clinical relevance as the largest randomized controlled trial thus far to evaluate the effect of protein restriction and/or BP control on kidney progression.

When the null hypothesis significance test fails to reach statistical significance, null findings are supported by a narrow CI (which must contain the null value since $P \geq \alpha$), illustrated by the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial study, a category "c" study, but one in which the study has great clinical relevance (Table 2; Figure 1, top right corner). In such situations, information is gained because a study might eliminate unnecessary/ineffective therapies and/or therapeutic maneuvers because, at best, the size of the effect is unlikely to be clinically significant. However, a null study with wide CI, illustrated by the MDRD study, a category "d" study (Table 2) can also be important even despite low result-based but on high process-based merit.

Exploring Outcomes beyond Statistical Issues

Although we have discussed type 2 errors at length, the concept of type 1 errors is also important. Recall that a type 1 error is the probability of rejecting the null hypothesis when the null hypothesis is in fact true. An example of a type 1 error in the literature might be the Crit-Line Intradialytic Monitoring Benefit study (20), a multicenter RCT in which 1227 hemodialysis patients were randomly assigned to intradialytic blood volume monitoring using Crit-Line and 216 to conventional monitoring for 6 mo to test the hypothesis of whether Crit-Line guided monitoring would mitigate the hospitalization rates in hemodialysis patients. There were 120 non–access-related hospitalizations in the Crit-Line monitoring group and 81 non–access-related hospitalizations in the conventional monitoring group (unadjusted RR for non–access-related hospitalization, 1.49; 95% CI, 1.07 to 2.08; $P = 0.017$) in the Crit-Line monitoring group compared with the conventional group. Mortality was 8.7% and 3.3% ($P = 0.021$) in the Crit-Line and conventional monitoring groups, respectively. Additionally, although an elaborate protocol was available to guide fluid management based on Crit-Line, investigators were not mandated to follow this protocol. In this trial, uncertain adherence to the protocol by the investigators makes it difficult to conclude that continuous blood volume monitoring was associated with higher complication rates. The internal validity of the trial may have been affected by noncompliance bias. Furthermore, as the authors themselves state, the atypically low hospitalization and mortality rates for the conventional monitoring group suggest that these findings may reflect bias rather than a spurious association by chance (type 1 error). Thus, this study, while a category "a" study (Table 2), could be regarded as inconclusive because of concerns regarding bias and internal validity, even although the P value reached conventional levels of significance. There is ambiguity as to whether or not there really is a difference in the Crit-Line and conventional monitoring groups. That is, the study had arguably high results-based merit and arguably low process-based merit (Figure 1, bottom right corner).

Is There a Publication Bias against Null Studies?

Journals flatly rejecting to publish studies because of the lack of statistically significant findings can be problematic (21–24). Indeed, outright rejection due solely to a lack of statistically significant results can stifle progress in a field by not contributing to the field's cumulative knowledge. Imagine a theory predicting that a certain treatment should have an effect above and beyond no treatment. Interested in evaluating the theory, a team of researchers test the hypothesis, and their results fail to reach statistical significance. Although their study was well designed and conducted, had sufficient statistical power to reject an effect of meaningful clinical significance, and had a narrow CI around the null value, their study was rejected outright because of the lack of statistical significance. Now imagine many teams of researchers conducting essentially this same study and each and every one having their manuscript rejected from publication because of a lack of statistically significant findings. Not only are resources wasted in reproducing the study again and again, but the estimated effect sizes and measures of variability are not reported to the research community for possible inclusion in meta-analyses because such studies are not being published.

In addition to journals not publishing studies without statistically significant findings, some data suggest that the authors themselves might withhold publication of results that fail to reach statistical significance (25). Publication bias can run the gamut from delaying publication to just saying no to publishing the study. The Cochrane Collaboration (21) compared time to publication for studies that showed positive findings and those that generated results that failed to reach statistical significance. Their data showed that trials containing statistically significant results were, on average, published sooner than those without statistically significant results. Stern and Simes (22), in a retrospective analysis of trials conducted between 1979 and 1988 at a single center (a university hospital in Australia), reached similar conclusions.

We want to make clear that failing to reject a null hypothesis, in and of itself, certainly does not imply that a study should be published, nor does rejecting a null hypothesis for that matter. Many other factors are necessary for a study to be worthy of publication. That being said, journals or guideline developers may appropriately resist publication or inclusion of studies without statistically significant findings for several reasons (26,27). For example, the trial may be poorly designed, erroneously implemented, or is simply "not so important" given the existing literature and current state of knowledge. On the other hand, by publishing well-designed studies that were appropriately implemented and provide insight into clinically and/or scientifically important findings, but did not produce statistically significant effects, a greater balance in evaluating the positive trials may be achieved.

Some have argued that all papers should be published and that meta-analysis should be used to draw conclusions based on large bodies of data (8). Interestingly, a journal for the publication of negative results, *Journal of Negative Results in BioMedicine*, is in place with the goal of "providing scientists and physicians with responsible and balanced information to support informed experimental and clinical decisions." Furthermore, meta-analysis is a very powerful tool that has not yet been fully appreciated, nor has it been used to its full potential (28–30). As meta-analysis becomes more widely used, it is likely that well-designed and conducted studies will be included in meta-analyses so that the file-drawer phenomenon does not lead to a biased sample of studies that are actually published (some of which are likely type 1 errors and others in which erroneous conclusions may have been reached because of bias). The file-drawer problem is where studies that fail to have statistically significant results are relegated to the file-drawer instead of being published (31).

The Importance of Publishing Null Studies

Clinical trialists generally try to design studies that reject the null hypothesis, *i.e.*, to produce "positive results." Unquestionably, this is the general goal because readers find "positive study results" informative and such results are likely to have impact. However, studies where the null hypothesis is not rejected can be important if the conditions outlined above are met; indeed, these studies should not be unilaterally excluded from consideration. Journal editors and reviewers should require authors to report effect sizes of interest as well as their corresponding 95% CIs and a clear discussion of the rationale of the sample size chosen, whether it be from the power analytic approach, the accuracy in parameter estimation approach, or based on some other rationale. As Greenwald (32) eloquently recommends to journal editors: "base publication decisions on criteria of importance and methodological soundness uninfluenced by whether a result supports or rejects a null hypothesis."

Acknowledgments

The authors thank Keke Lai, BS (Inquiry Methodology Program Indiana University, Bloomington, IN) for insights and comments.

Disclosures

None.

References

1. Motulsky H: Intrepreting nonsignificant P values. In: *Intuitive Biostatistics*, Oxford, Oxford University Press, 1995
2. Altman DG, Bland JM: Absence of evidence is not evidence of absence. *BMJ* 311: 485, 1995
3. Hartung J, Cottrell JE, Giffin JP: Absence of evidence is not evidence of absence. *Anesthesiology* 58: 298–300, 1983
4. Steiger JH, Fouladi RT: Noncentrality interval estimation and the evaluation of statistical models. In: *What If There Were No Significance Tests*? edited by Harlow LL, Mulaik SA, Steiger JH, Hillsdale, NJ, Lawrence Erlbaum, 1997
5. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR: The importance of β, the type II error and sample size in the design and interpretation of the randomized clinical trial: survey of 71 negative trials. *N Engl J Med* 299: 690–694, 1978
6. Alderson P: Absence of evidence is not evidence of absence. *BMJ* 328: 476–477, 2004
7. Maxwell SE, Kelley K, Rausch JR: Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 59: 537–563, 2008
8. Aberson C: Interpreting null results: improving presentation and conclusions with confidence intervals. JASNH 93: 36–42, 2002
9. Cashen LH, Geiger SW: Statistical power and the testing of null hypotheses: a review of contemporary management research and recommendations for future studies. *Organizational Res Methods* 7: 151–167, 2004
10. Goodman SN, Berlin JA: The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 121: 200–206, 1994 [erratum in *Ann Intern Med* 122:478, 1995]
11. Cohen J: *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed, Hillsdale, NJ, Lawrence Erlbaum, 1988
12. Kelley K: Sample size planning for the coefficient of variation: accuracy in parameter estimation via narrow confidence intervals. *Behav Res Methods* 39: 755–766, 2007
13. Kelley K, Maxwell SE, Rausch JR: Obtaining power or obtaining precision: delineating methods of sample-size planning. *Eval Health Prof* 26: 258–287, 2003
14. Drueke TB, Locatelli F, Clyne N, Eckardt KU, Macdougall IC, Tsakiris D, Burger HU, Scherhag A: CREATE Investigators: normalization of hemoglobin level in patients with chronic kidney disease and anemia. *N Engl J Med* 355: 2071–2084, 2006
15. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA* 288: 2981–2997, 2002 [erratum in *JAMA* 289:178, 2003; *JAMA* 291:2196, 2004]
16. Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, Schwab SJ, Goodkin DA: The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. *N Engl J Med* 339: 584–590, 1998
17. Macdougall IC, Ritz E: The Normal Haematocrit Trial in dialysis patients with cardiac disease: are we any the less confused about target haemoglobin? *Nephrol Dial Transplant* 13: 3030–3033, 1998
18. Ginzler EM, Dooley MA, Aranow C, Kim MY, Buyon J, Merrill JT, Petri M, Gilkeson GS, Wallace DJ, Weisman MH, Appel GB: Mycophenolate mofetil or intravenous cyclophosphamide for lupus nephritis. *N Engl J Med* 353: 2219–2228, 2005
19. Klahr S, Levey AS, Beck GJ, Caggiula AW, Hunsicker L, Kusek JW, Striker G: The effects of dietary protein restriction and blood pressure control on the progression of chronic renal disease. *N Engl J Med* 330: 877–884, 1994
20. Reddan DN, Szczech LA, Hasselblad V, Lowrie EG, Lindsay RM, Himmelfarb J, Toto RD, Stivelman J, Winchester JF, Zillman LA, Califf RM, Owen WF Jr: Intradialytic blood volume monitoring in ambulatory hemodialysis patients: a randomized trial. *J Am Soc Nephrol* 16: 2162–2169, 2005
21. Hopewell S, Clarke M, Stewart L, Tierney J: Time to publication for results of clinical trials. *Cochrane Database Syst Rev* 2: MR000011, 2007
22. Stern JM, Simes RJ: Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 315: 640–645, 1997
23. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR: Publication bias in clinical research. *Lancet* 337: 867–872, 1991
24. Dickersin K, Min YI: Ann Publication bias: the problemthat won't go away. *NY Acad Sci* 703: 135–146, 1993; discussion 146–148
25. Dickersin K, Min YI, Meinert CL: Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 267: 374–378, 1992
26. Ionannides J: Why most published research findings are false. *PloS Med* 2: e124, 2005
27. Cleophas RC, Cleohas TJ: Is selective reporting of clinical

research unethical as well as unscientific? *Int J Clin Pharmacol Ther* 37: 1–7, 1999

28. Hunter JE, Schmidt FL. *Methods of Meta-analysis*: *Correcting Error and Bias in Research Findings*, Thousand Oaks, CA, Sage, 2004
29. Hedges LV, Olkin I: *Statistical Methods for Meta-analysis*, Orlando, FL, Academic Press, 1985
30. Sutton AJ, Jones DR, Abrams KR, Sheldon TA, Song F: *Methods for Meta-analysis in Medical Research*, London, John Wiley, 2000
31. Rosenthal R: The file drawer problem and tolerance for null studies. *Psychol Bull* 86: 638–641, 1979
32. Greenwald AG: Consequences of prejudice against the null hypothesis. In: *A Handbook for Data Analysis in the Behavioural Sciences*, edited by Keren G, Lewis C, Hillsdale, NJ, Lawrence Erlbaum, 1993, pp 419–448