JACC: CARDIOVASCULAR INTERVENTIONS
VOL. 1, NO. 3, 2008

ISSN 1936-8798/08/$34.00
PUBLISHED BY ELSEVIER INC.
DOI: 10.1016/j.jcin.2008.01.008

STATE-OF-THE-ART PAPERS

Randomized Clinical Trials and Observational Studies

Guidelines for Assessing Respective Strengths and Limitations

Edward L. Hannan, PhD, MS, MS, FACC

Rensselaer, New York

The 2 primary types of studies that are used to test new drugs or procedures or compare competing drugs or types of procedures are randomized clinical trials (RCTs) and observational studies (OS). Although it would appear that RCTs always trump OS because they eliminate selection bias, there are many possible limitations to both types of studies, and these limitations must be carefully assessed when comparing the results of RCTs and OS. This state-of-the art review describes these limitations and discusses how to assess the validity of RCTs and OS that yield different conclusions regarding the relative merit of competing treatments/interventions. (J Am Coll Cardiol Intv 2008;1:211–7) © 2008 by the American College of Cardiology Foundation

The 2 primary means by which alternative medical or surgical treatments are assessed is through the use of randomized controlled trials (RCTs) or observational studies (OS). The RCTs were first introduced when streptomycin was evaluated as a treatment for tuberculosis (1,2). In RCTs, participants are randomly assigned to a treatment or control group (or to multiple treatment groups) so as to reduce bias by making the groups as equal as possible with respect to all patient characteristics that may have an impact on outcomes. Thus, in theory, the only difference between the groups is the treatment assignment and any differences that are identified. In contrast, OS do not randomize treatment but "observe" differences in outcomes that occur after treatment decisions have been made without regard to ensuring that patients in different treatment arms have similar characteristics related to outcomes.

Evidence-based medicine classifies different types of studies on the basis of research design as the criterion for hierarchical rankings (2–4). Table 1 presents the American College of Cardiology/American Heart Association (ACC/AHA) levels of evidence for treatment recommendations (4). As noted in the table, the highest level of evidence (Level A) is accorded to studies with "data derived from multiple randomized clinical trials or meta-analyses" and the second highest level (Level B) is assigned to studies with "data derived from a single randomized trial, or non-randomized studies" (4).

It is assumed that a study design at a higher level in the hierarchy is methodologically superior to one at a lower level and that studies at the same level are equivalent, but this ranking system is far too simplistic given the many design characteristics that comprise a given study.

Other hierarchies that are more comprehensive than the American College of Cardiology/American Heart Association Levels of Evidence include phrases such as "properly randomized, controlled trial" (5), "well-designed controlled trials without randomization" (5), and "RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise)" (6). However, even in these more detailed hierarchies, it is not clear how to revise the hierarchy if a trial is not "properly randomized" or "well designed." Also, the randomization and the quality of the design are not necessarily best described in a dichotomous manner; rather there are many factors to consider

From the University at Albany School of Public Health, Rensselaer, New York.

Manuscript received October 24, 2007; revised manuscript received December 17, 2007, accepted January 10, 2008.

in designing a study, and it would appear that the boundaries of any hierarchy may be blurred by the quality of a study's design.

The purpose of this communication is to explore the pros and cons of RCTs and OS in general, but more importantly to propose some criteria for evaluating the quality of RCTs and OS that will provide some insight regarding which option is appropriate for a given problem, and how to assess the validity of RCTs and OS that yield different conclusions regarding the relative merit of competing treatments/interventions.

Comparing Evidence from RCTs and OS

First, it is informative to review the results of some relatively recent studies that have used up-to-date meta-analytic techniques to combine the outcomes of similar OS and similar RCTs that have studied the same problem. Benson and Hartz (7) examined OS published between 1985 and 1998 that were aimed at comparing 2 or more treatments or interventions for the same condition. The Abridged Index Medicus and Cochrane databases were used to find all RCTs and OS that compared the same treatments for those conditions. Benson and Hartz identified a total of 136 different reports on 19 diverse treatments, including 7 cardiologic treatments: nifedipine versus control in patients with coronary artery disease; coronary artery bypass graft (CABG) surgery versus percutaneous transluminal coronary angioplasty (PTCA) in diabetic patients; CABG surgery versus PTCA in patients at high risk; CABG surgery versus PTCA in patients at low risk; CABG surgery versus medical treatment in the CASS (Coronary Artery Surgery Study) trial; CABG surgery versus medical treatment in Duke study patients; and the use of beta-blockers versus placebo (7). Their conclusions were that in only 2 of the 19 treatments did the combined magnitude of the effect in the OS lie outside the 95% confidence interval (CI) for the combined magnitude of the effect in the RCTs. One of the 2 discrepant areas was in a cardiologic treatment: the comparison of CABG and PTCA for patients at low risk (7). Benson and Hartz concluded that "we find little evidence that estimates of treatment effects in OS reported after 1984 are either consistently larger than or qualitatively different than those obtained in randomized, controlled trials" (7).

Abbreviations and Acronyms

BMS = bare-metal stent(s)
CABG = coronary artery bypass graft
CI = confidence interval
DES = drug-eluting stent(s)
HRT = hormone replacement therapy
OR = odds ratio
OS = observational study/studies
PES = paclitaxel-eluting stent(s)
PTCA = percutaneous translumincal coronary angioplasty
RCT = randomized controlled trial(s)
RR = relative risk

Table 1. American College of Cardiology/American Heart Association Levels of Evidence for Research Studies

Level of Evidence: A	Data derived from multiple randomized clinical trials or meta-analyses
Level of Evidence: B	Data derived from a single randomized trial, or unrandomized studies
Level of Evidence: C	Consensus opinion of experts, case studies, or standard of care

Concato et al. (2) conducted a search for meta-analyses that compared outcomes of RCTs and OS. Five major medical journals were searched from 1991 to 1995, and summary estimates and confidence intervals were computed from a total of 99 reports across 5 topic areas. These areas included the treatment of hypertension and stroke, the treatment of hypertension and coronary heart disease, and cholesterol level and death due to trauma. The 2 other areas were breast cancer and tuberculosis.

They concluded that the mean results from the OS were remarkably similar to the results from the RCTs. For example, for treatment of hypertension and stroke, the summary estimates were relative risk (RR) 0.58 (95% CI 0.50 to 0.67) for RCTs and odds ratio (OR) 0.62 (95% CI 0.60 to 0.65) for OS. For treatment of hypertension and coronary heart disease they were RR 0.86 (95% CI 0.78 to 0.96) and OR 0.77 (95% CI 0.75 to 0.80), respectively. Concato et al. (2) concluded that the "results of well-designed studies. . .do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic."

A recent study by Ioannidis et al. (8) compared the results of evaluations of 45 medical treatments based on 240 randomized trials and 168 nonrandomized studies to judge the consistency of results between the 2 types of studies. They found that there was very good correlation between the summary odds ratios ($R = 0.75$, $p < 0.001$), but that the nonrandomized studies were more likely to show larger treatment effects (28 vs. 11, $p = 0.009$). They also found that there was frequent heterogeneity among study results among randomized studies and among nonrandomized studies (8).

Complementarity of RCTs and OS

Thus, for the most part, RCTs and OS do arrive at the same conclusions. Furthermore, RCTs and OS can be used synergistically to obtain more and better information about the relative merits of alternative interventions/treatments. For example, OS can be used to:

- Test the external validity of RCTs by expanding the settings to a more representative population (9);
- Formulate hypotheses for RCTs to test (9);

- Identify structures, processes, and outcomes to study (10);
- Help establish the appropriate sample size for an RCT (10); and
- Examine patient subsets to determine precisely which patients benefit from each alternative intervention

Reasons Why RCTs and OS May Arrive at Different Conclusions

However, it is not always true that RCTs and OS arrive at the same conclusions and, consequently, it is important to ask what should be done when they do not agree. First, as noted in Table 1, the ACC/AHA require multiple RCTs or a meta-analysis for their highest level of evidence and multiple OS (or 1 RCT) for their second-highest level. Thus, if 1 RCT differs from multiple OS, then there is a pressing need to examine the studies more closely. The sections to follow list reasons why the 2 types of studies may arrive at different conclusions and of potential threats to validity of observational databases and RCTs.

Selection bias. Perhaps the most serious shortcoming of OS is selection bias, whereby because of the absence of randomization, there may be large observed and unobserved differences in patient characteristics between the treatment and control (or between 2 or more treatment) groups (10). These differences can lead to biased estimates of the treatment effects when one or more of the patient characteristics for which there are differences are related to the outcomes being measured (10). These factors are referred to as confounders. RCTs were developed for the purpose of eliminating this bias.

As an example of this problem, a meta-analysis of all 25 OS published through 1997 found that the RRs for coronary heart disease for patients who ever used hormone replacement therapy (HRT) relative to patients who never used it were 0.70 (95% CI 0.65 to 0.75) for estrogen only and 0.66 (95% CI 0.53 to 0.84) for the combination of estrogen and progestin (11). However, in 2002, the Women's Health Initiative study, an RCT with 16,000 postmenopausal women that was planned for an 8-year follow-up, was stopped because of an increase in the risk of breast cancer and heart disease in women treated with HRT. This finding led to recommendations that ". . .post-menopausal women who are considering estrogen or estrogen with progestin treatments should discuss with their health care providers whether the benefits outweigh the risks" (12). It is hypothesized that the reason for this discrepancy was a selection bias in the OS, whereby women taking HRT were more likely to be in higher socioeconomic groups who have had better access to preventive health care their entire life and were therefore less likely to experience adverse health outcomes (12).

There are several ways in which selection bias related to known factors can be controlled for and reduced, although not entirely eliminated, including risk adjustment through regression or analysis of variance methods (13–15), propensity analysis (13,16,17), and instrumental variables (13,18).

Risk-adjustment generally involves the development of a statistical model (when the outcome is binary, usually either a logistic regression model for short-term outcomes or a proportional hazards model for longer term outcomes) with the outcome as a dependent variable. The type of treatment is used as an independent variable along with the control variables, which are typically patient risk factors suspected or known to be related to adverse outcomes. Then the impact of treatment type (OR for short-term outcomes or hazard ratio for longer-term outcomes) is a byproduct of the statistical model that measures the impact of treatment after adjusting for the control variables.

Propensity analysis is a method developed to match patients in an observational study as well as possible with regard to characteristics that are associated with the choice of treatment. Typically, this is done by developing a logistic regression model that has choice of treatment as a binary dependent variable and the characteristics that are potentially associated with treatment choice as the independent variables. Then, the probability of one of the treatments (say Treatment A) being chosen can be calculated from the model as a function of the characteristics, and pairs of patients receiving each treatment and having identical or similar probabilities of Treatment A being chosen can be identified. Another way of controlling for propensity to receive one of the treatments is to use the propensity score as an additional independent variable in the risk-adjustment model in addition to the patient characteristics. However, if the variables used to predict the propensity score are identical or nearly identical to the other variables in the model, this variation of the method is not effective.

Instrumental variables are another method for controlling for selection bias that require the use of a variable that is related to treatment choice but not to outcomes. This approach is not as popular as propensity analysis, at least in part because it is sometimes difficult to identify an instrumental variable.

Perhaps the greatest threat to selection bias is related to unobserved differences in patient characteristics, i.e., patient characteristics that are not contained in the observational database (particularly data collected for clinical purposes). These characteristics cannot be controlled for nor can their impact be measured using the methods described previously. Therefore, if unobserved characteristics are significant predictors of outcome and if they are unbalanced with respect to alternative treatments, then the potential for significant bias exists in OS.

The presence of unobserved differences in patient characteristics should be minimized if possible by designing

JACC: CARDIOVASCULAR INTERVENTIONS, VOL. 1, NO. 3, 2008
JUNE 2008:211–7

observational databases so as to include as many patient characteristics that are thought to impact outcomes as possible. Also, if there are patient characteristics that are associated with contraindication of one of the treatments being considered, these contraindications should be included in the database if possible. For example, if drug-eluting stents (DES) are being compared with bare-metal stents (BMS) in a time period in which DES are unavailable in longer lengths, then either the lesion length could be included in the database or the fact that DES was contraindicated because of the lesion length could be included as a data element. Thus, either lesion length could be controlled for or patients with long lesions could be excluded from the analyses.

The danger of selection bias is exacerbated when the observational database is an administrative database rather than a clinical database. An administrative database is a database that is developed for purposes of reimbursement or planning as opposed to being created to compare treatments/interventions or to evaluate quality of care. The major drawbacks of administrative databases in evaluating quality/comparing risk-adjusted outcomes for interventions are the limited ability to distinguish between complications of care (adverse outcomes) and pre-existing conditions (risk factors in the risk-adjustment process), inability to specify clinical definitions for risk factors (e.g., forced reliance on International Classification of Diseases-9 codes), and limitations on the number of risk factors that are coded for each patient (19–24). Despite these limitations, administrative databases do have great potential for evaluating care, particularly if a limited number of clinical data elements can be added to them (19,23). Another advantage of administrative databases is that because they were not created for a priori-defined research purposes, there is no opportunity for interviewer bias or ascertainment bias.

The presence of an important selection bias constitutes low internal validity of a study (ability of inferences from the study to represent cause-effect relationships). Well-done RCTs are superior to OS because they eliminate selection bias. However, there are many lower quality RCTs that suffer from deficits in external validity (the extent to which the results can be generalized beyond the sample). The following are some potential problems that relate to external validity.

Generalizability. One reason why an RCT and an observational study on the same competing interventions may arrive at different conclusions is that they frequently apply to different patients. Randomized controlled trials have specific inclusion and exclusion criteria that are often quite restrictive, whereas OS usually apply to a much broader population and are frequently even population-based. There is evidence that RCT populations usually don't mirror the age, gender, and race distribution of the target patient population (25–29). In general, they tend to be less sick, younger, better educated, and of higher socioeconomic status. This also means that in RCTs, patients are more likely to be adherent. This may tend to overstate the effect of a new treatment had it been introduced in the entire target population. In the case of RCTs that involve procedure-based interventions, informed consent can be more difficult to obtain than it is in drug trials. As an example, it has been estimated that as few as 2% to 5% of the patients screened were randomized in the early PTCA versus CABG trials (10). Thus, extrapolation to the entire population may be unwise. Consequently, there is a possibility that if the OS had been restricted to the same set of exclusions and inclusions as the RCT, its results may have been much more similar. On the other hand, the results for all patients in the observational study are of great interest because they reflect actual practice patterns and because they enable researchers to conduct subset analyses that will speak to precisely which patients benefit from each treatment/intervention.

Another threat to the generalizability of an RCT is that the clinicians or providers in the study are not representative of the population of clinicians or hospitals who would ultimately be using the procedure/treatment. As an example, off-pump CABG surgery is more demanding than traditional on-pump surgery and it is likely that outcomes are more variable as a function of the skill or experience of the surgeon performing the surgery. If an RCT comparing off-pump and on-pump surgery is conducted by a group of surgeons with unusual skill/experience in off-pump procedures, the conclusion of the RCT that off-pump surgery is superior to on-pump surgery may not translate to the general population of CABG surgery patients. For example, van Dijk et al. (30) found 5-year mortality rates of 8.5% for off-pump patients and 6.5% for on-pump patients in a RCT performed in the Netherlands, whereas Hannan et al. (31) found 3-year rates of 10.6% for off-pump and 9.9% for on-pump patients in a population-based study in New York. Reasons for the greater rates in New York may include lower-risk patients in the RCT, but also the fact that surgeons in the RCT had already traversed the longer learning curve required for the demands of off-pump surgery.

Inadequate statistical power. Because of the cost of RCTs, the amount of time it takes to conduct them, the difficulty recruiting subjects, and the fact that because they generally expect only small to modest differences in outcomes (otherwise they would be unethical), they are frequently underpowered to detect important differences in outcomes. This can lead to erroneous conclusions, generally false negatives, i.e., that there are no significant differences in treatments when a larger sample size would have uncovered significant differences.

To achieve adequate statistical power, researchers can, and usually do conduct power analyses to determine the sample size necessary to identify meaningful clinical differences between treatments. However, because of cost and other considerations, compromises are frequently made.

Also, power analyses are predicated on assumptions about adverse outcome rates, which may prove to be inaccurate. One of the ways that RCTs attempt to combat the lack of sufficient statistical power is to use combined outcomes (e.g., major adverse cardiac events rather than mortality) to enhance the adverse outcome rate to compensate for inadequate sample size. However, a major drawback of this substitution is that outcomes with far different levels of severity are combined, and the results are frequently driven by less important, and sometimes even subjective, outcomes with greater incidence rates.

For example, Freemantle et al. (32) report that in a systematic review of the use of composite end points that included mortality in clinical trials published in 9 top medical journals between 1997 and 2001, among the 79 trials in which the composite end point yielded a statistically significant result, the mortality end point by itself was only significant in 19 trials (24%). Freemantle et al. (32) concluded that ". . .reporting of composite outcomes is generally inadequate, implying that the results apply to the individual components of the composite outcome rather than only to the overall composite." In an accompanying editorial, Lauer and Topol (33) concluded that "when composite end points are used, the individual components must be appropriately chosen, objectively measured in an unbiased manner, and individually reported."

Although this is excellent advice to follow, there remains the danger that important components such as mortality are reported separately and found to be nonsignificant, but seemingly only because of inadequate statistical power. As an example, in the PASSION (Paclitaxel-eluting Stent vs. Conventional Stent in Myocardial Infarction with ST-segment Elevation) trial, which compared outcomes for patients with ST-segment elevation myocardial infarction who were treated with BMS and paclitaxel-eluting stents (PES), Laarman et al. (34) assigned 619 ST-segment elevation myocardial infarction patients to PES and BMS. The primary end point of the study was a composite of death from cardiac causes, recurrent myocardial infarction, or target lesion revascularization at 1 year. Findings were that there was no significance in the rates of serious adverse events (8.8% vs. 12.8% in favor of PES, adjusted RR 0.63, 95% CI 0.37 to 1.07, p = 0.092) (34). Nevertheless, it appears that the magnitude of these differences is clinically important. In fact, with roughly 600 patients in each group, the statistical power to detect a difference between 8.8% and 12.8% is only 72%. This power increases to a respectable 89% if the samples are of size 1,000 instead of 600.

In conclusion, combined end points and low statistical power can lead to misleading conclusions, but these problems can be overcome if there are sufficient resources available to obtain adequate sample sizes. However, in my view many RCTs have not had adequate sample sizes for testing what should have been the primary end point.

Follow-up and approach to treating patients. Because RCTs generally have strictly defined follow-up criteria whereas follow-ups in OS are driven by variable physician-scheduled appointments and decisions of patients to seek care based on symptoms, different information may be available in the 2 types of studies. These differences can lead to biases in each study. For example, the absence of follow-ups in OS may result in uncaptured complications that resulted in outpatient visits that are not contained in the observational database.

However, it is also possible that mandated follow-ups in RCTs may bias outcomes in favor of one of the treatments. For instance, in the COURAGE (Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation) trial that compared optimal medical therapy with and without PCI for patients with stable coronary disease, compliance rates with medical therapy in the medically treated cohort at 5 years of follow-up were 94% for aspirin, 93% for statins and 86% for beta-blockers, respectively (35).

However, the CRUSADE (Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes with Early Implementation of the American College of Cardiology/American Heart Association Guidelines) registry found only 46% of patients were compliant with beta-blockers alone, and only 21% were compliant with beta-blockers, aspirin, and lipid-lowering therapy together (36,37). Thus, the rates reported in the COURAGE trial are unlikely to be reproducible in real-world settings. In addition, medical noncompliance has been shown to be associated with higher mortality in late follow-up (36,38). Hence, it is likely that the case-management nature of the follow-up in the COURAGE trial resulted in considerably better outcomes for medically treated patients than would occur in typical settings.

Criteria for Evaluating Quality of RCTs and OS

In view of the possible threats to validity in both RCTs and OS, the following are some questions to ask when evaluating the quality of a study:

Quality of the database. Does the database contain all of the characteristics/variables known to be necessary to obtain valid conclusions? Are the variables collected and measured in a well-defined clinically meaningful manner? If it is an observational database, does it contain the patient risk factors known to be significant predictors of the outcomes being tracked and studied? Is unmeasured confounding/selection bias a significant threat to the findings of the study?

Patients. Are the patients in the study the right patients to test the study hypothesis? If the study is an observational database, is there enough information about patients to make appropriate exclusions? If the study is an RCT, are the exclusion and inclusion criteria broad enough to inform broad-based treatment decisions made on the basis of the study findings?

Outcomes. Are the outcomes used in the study meaningful ones? Are there combined end points that mix important outcomes (e.g., mortality) with relatively unimportant or subjective outcomes. Are important outcomes included individually?

Size and generality of the database. Is the database large enough to yield the statistical power needed to identify clinically meaningful differences in the important outcomes? "Clinically meaningful differences" should be defined in advance of the study. Is the sample of patients generalizable to other settings? Is there just a single site or very few sites that may not be representative of outcomes at other sites because of special circumstances such as physician quality, hospital quality or exceptional resources?

Analysis strategy. If the study is based on an observational database, are differences in patient risk factors between treatments being controlled for adequately using a combination of multivariable adjustment and a method for testing selection bias such as propensity analysis?

Follow-up. Is the follow-up period long enough to capture the outcomes being evaluated? Is the follow-up process complete and does it mirror real-world practice? Is the study compromised by loss to follow-up?

Conclusions

The aforementioned criteria (and perhaps others I have inadvertently omitted) are the most important determinants of whether a database and the methodology for analyzing it are adequate for obtaining valid conclusions, or whether a given database/analysis plan is superior to another one, regardless of whether the database is an RCT or an observational database. The design and ultimate conduct of the study is the principal criterion to consider, not the type of study per se (19,20).

Reprint requests and correspondence: Dr. Edward L. Hannan, School of Public Health, State University of New York, Department of Health Policy, Management and Behavior, SUNY University at Albany, One University Place, Rensselaer, New York 12144-3456. E-mail: elh03@health.state.ny.us.

REFERENCES

1. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. BMJ 1948;2:769–82.
2. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92.
3. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992;268:2420–5.
4. American Heart Association. Methodology Manual for ACC/AHA Guideline Writing Committees. June 2006. Available at: http://www.americanheart.org/presenter.jhtml?identifier=3039684. Accessed April 3, 2008.
5. Preventive Services Task Force. Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force. 2nd edition. Baltimore, MD: Williams and Wilkins; 1996.
6. Guyatt G, Gutterman D, Baumann MH, et al. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force. Chest 2006;129:174–81.
7. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. N Engl J Med 2000;342:1878–86.
8. Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821–30.
9. Stables RH. Observational research in the evidence based environment: eclipsed by the randomised controlled trials? Heart 2002;87:101–2.
10. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–8.
11. Barrett-Connor E, Grady D. Hormone replacement therapy, heart disease, and other considerations. Annu Rev Public Health 1998;19:55–72.
12. U.S. Department of Health and Human Services. Facts About Menopausal Hormone Therapy. Available at: http://www.nhlbi.nih.gov/health/women/pht_facts.pdf. Accessed September 15, 2007.
13. D'Agostino RB Jr., D'Agostino RB Sr. Estimating treatment effects using observational data. JAMA 2007;297:314–6.
14. Hannan EL, Racz MJ, Walford G, et al. Long-term outcomes for coronary artery bypass graft surgery vs. stent implantation. N Engl J Med 2005;352:2174–83.
15. Hannan EL, Racz M, Holmes DR, et al. The impact of completeness of revascularization on long-term outcomes in the stent era. Circulation 2006;113:2406–12.
16. Rosenbaum PR, Rubin DR. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–24.
17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
18. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA 1994;272:859–66.
19. Hannan EL, Kilburn H Jr., Lindsey ML, et al. Clinical versus administrative data bases for CABG surgery: does it matter? Med Care 1992;30:892–907.
20. Jollis JG, Ancukiewicz M, DeLong ER, et al. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med 1993;119:844–50.
21. Iezzoni LI, Ash AS, Coffman GA, et al. Predicting in-hospital mortality: a comparison of severity measurement approaches. Med Care 1992;30:347–59.
22. Iezzoni LI, Ash AS, Shwartz M, et al. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. Ann Intern Med 1995;123:763–70.
23. Hannan EL, Racz MJ, Jollis JG, et al. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? Health Serv Res 1997;31:659–78.
24. Pine M, Jordan HS, Elixhauser A, et al. Enhancement of claims data to improve risk adjustment of hospital mortality. JAMA 2007;297:71–6.
25. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999;319:312–5.
26. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials-race, sex and age-based disparities. JAMA 2004;291:2720–6.
27. Sorensen HT, Lash T, Rodman KJ. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies. Hepatology 2006;44:1075–82.

28. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women and minorities in heart failure clinical trials. Arch Intern Med 2002;162:1682–8.
29. Buring JE. Women in clinical trials: a portfolio for success. N Engl J Med 2000;343:505–6.
30. van Dijk D, Spoor M, Nathoe HM, et al. Cognitive and cardiac outcomes 5 years after off-pump vs. on-pump coronary artery bypass graft surgery. JAMA 2007;297:701–8.
31. Hannan EL, Wu C, Smith CR, et al. Off-pump vs. on-pump CABG surgery: differences in short-term outcomes and in long-term mortality and need for subsequent revascularization. Circulation 2007;116: 1145–52.
32. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials. JAMA 2003;289:2554–9.
33. Lauer MS, Topol EJ. Clinical trials-multiple treatments, multiple endpoints, and multiple lessons. JAMA 2003;289:2575–7.
34. Laarman GJ, Suttorp MJ, Dirksen MT et al. Paclitaxel-eluting versus uncoated stents in primary percutaneous coronary interventions. N Engl J Med 2006;355:1105–13.
35. Boden WE, Orourke RA, Teo KK, et al. Optimal medical therapy with or without PCI for stable coronary disease. N Engl J Med 2007;35: 1503–16.
36. Kereiakes DJ, Teirstein PS, Sarembock IJ, et al. The truth and consequences of the COURAGE trial. J Am Coll Cardiol 2007;50: 1598–603.
37. Mehta RH, Roe MT, Chen AY, et al. Changing practice for non ST-segment elevation acute coronary syndromes: trends from the CRUSADE quality improvement initiative (abstr). Circulation 2005; 112:II793.
38. Gallagher EJ, Viscoli CM, Horwitz RI. The relationship of treatment adherence to the risk of death after myocardial infarction in women. JAMA 1993;270:742–4.