ALLIANCE for ACADEMIC INTERNAL MEDICINE

AAIM Perspectives

AAIM is the largest academically focused specialty organization representing departments of internal medicine at medical schools and teaching hospitals in the United States and Canada. As a consortium of five organizations, AAIM represents department chairs and chiefs; clerkship, residency, and fellowship program directors; division chiefs; and academic and business administrators as well as other faculty and staff in departments of internal medicine and their divisions.

Challenges of Assessing Therapeutic or Diagnostic Outcomes with Observational Data

William S. Weintraub, MD,[a] Robert W. Yeh, MD, MSc, MBA[b]

[a]MedStar Washington Hospital Center, Northwest DC; [b]Beth Israel Deaconess Medical Center, Boston, Mass.

KEYWORDS: Epidemiology; Observational data; Outcomes research; Selection bias

Medical care is replete with decisions on diagnostic approaches and therapeutic strategies. To establish the evidence that can help clinicians decide on the correct choice, patients are sampled from the larger population under consideration, studied in detail, and results in these patients are then applied more broadly. Two approaches are observational studies and randomized trials. Publications resulting from observational studies will often offer considerable detail on their methods and then in the limitations section offer a statement to the effect that "While residual selection bias cannot be excluded, we controlled for all measured differences between the treatment groups." What does it mean; how important is it, in general and in any one study; and are randomized trials necessary to overcome this particular problem?

In randomized trials, patients drawn from the larger population are assigned to one therapy or the other by a process of random selection. For a randomized trial to proceed properly, there should be equipoise between the study arms, that is, patients and clinicians do not favor one choice over the other. The randomized trial has been the gold standard approach for the singular reason that it can, at least in principle, eliminate the bias that might be created by the nonrandom selection of treatment by clinicians and patients, that is, "treatment selection bias." Randomization results in the subjects in each arm having similar characteristics, both characteristics that are measured and that are unmeasured. Another type of selection bias can occur in randomized trials, in which the patients selected for a trial are not representative of patients being considered for a therapy, that is, the trial results cannot be generalized to the broader patient population. In all studies randomized or not, selection of appropriate patient populations to resolve clinical questions is crucial. Randomized trials are also expensive to mount, can become outdated, and then may not be repeated due to either lack of resources or lack of equipoise.

Due to the limitations of randomized trials, investigators have considered various nonrandomized approaches. While many variations exist, they generally fall into 2 categories, case-control and cohort studies. In case-control studies, patients with and without an outcome of interest are compared for a prior exposure. In a cohort study, patients are followed from a point of inception, and then individuals with and without exposure can be compared for the incidence of an outcome

Funding: Funded in part by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number U54-GM104941 (Principal Investigator: Binder-Macleod).

Conflicts of Interest: There are no conflicts of interest.

Authorship: Both authors had access to the data and a role in writing the manuscript.

Requests for reprints should be addressed to William S. Weintraub, MD, MedStar Heart & Vascular Institute Suite 4B1, MedStar Washington Hospital Center, 110 Irving Street Northwest, DC 20010.

E-mail address: william.s.weintraub@medstar.net

https://doi.org/10.1016/j.amjmed.2017.10.027

of interest. In both case control and cohort studies, there can be difficulties with exposure due to variable adherence or crossover from one form of therapy to the other.

The major limitation with an observational study compared with a randomized trial is the influence of "treatment selection bias."[1] Treatment selection bias can occur when the therapeutic selection is influenced by patient characteristics, including severity and acuteness of illness and comorbidity. When such variables also are associated with better or worse outcomes, they are called confounders. If unaccounted for, confounders will bias the result of observational studies, potentially resulting in erroneous conclusions. The literature is replete with observational studies that showed different results from randomized trials, likely due to confounding.[2,3] Statistical methods can be used to reduce the bias due to confounding where the confounding variables are measured. However, unmeasured confounders may make it difficult to assess whether a difference (or lack thereof) between the treatment arms is due to therapeutic effect or to residual confounding. Missing or misclassification of covariates (patient characteristics) is a particular problem in observational studies; in contrast to randomized trials, these data are necessary to correct for potential confounding. In all studies, randomized or not, correct classification of outcomes and as complete follow-up as possible are crucial.

PERSPECTIVES VIEWPOINTS

- Observational assessments of therapeutic or diagnostic options are subject to treatment selection bias.
- Statistical methods can reduce treatment selection bias but cannot account for unmeasured confounders.
- Simulation modeling can help assess whether there is residual selection bias.
- Sample size cannot overcome bias.
- Data are most believable when randomized clinical trial data and observational point in the same direction.

STATISTICAL METHODS TO REDUCE TREATMENT SELECTION BIAS

Various statistical approaches have been proposed to reduce treatment selection bias. The most common approaches employ variations of multivariable analysis. The most straightforward is to use logistic regression or Cox modeling, in which the treatment is included as a covariate along with measured potential confounders. Another series of approaches use propensity analysis.[4] This approach starts with identifying a set of variables that are related to the propensity to choose one form of therapy over the other. Using a logistic regression model, the probability of receiving one form of therapy vs the other (ie, the propensity score) may be calculated for each patient. Propensity scores can be used in several different ways, including creating matched groups, stratifying outcomes, or by reweighting of samples based on the inverse probability of receiving the treatment received.[5] All of these methods for overcoming treatment selection bias are limited to accounting for the bias induced by measured variables. None can account for unmeasured confounding variables. A recent study by Elze et al[6] found little, if any, advantage to propensity score approaches over multivariable analysis in the ability to reduce treatment selection bias.

An alternative approach to propensity analysis and other multivariate techniques is to use an instrumental variable.[7] An instrumental variable should cleanly separate the groups, with the distribution of variables the same in the resulting groups, except for the treatment variable, which should be quite different between the groups. The instrumental variable should not be associated with the outcome, apart from through its association with the treatment variable. The best instrumental variable is randomization, which cleanly separates the groups, but does not by itself predict outcome. An increasingly popular instrumental variable is Mendelian randomization in a genetic study.[8] For instance, the Mendelian randomization to familial hypercholesterolemia results in much higher risk of cardiovascular disease than individuals not randomized to familial hypercholesterolemia.[9] Attempts to develop other instrumental variables, such as geographic location or practice styles of physicians, have been used but may not fulfill the necessary assumptions for validity, because of differences that remain between the groups in critical variables, including unmeasured confounders.[10]

The effect of an unmeasured or a group of confounders may be evaluated by sensitivity analysis. In the method of Lin et al,[11] the observed difference in outcome is considered with the potential prevalence of the confounder in each arm. For any such pair of prevalences of the confounder in the 2 arms, the strength of the confounder, generally expressed as a hazard ratio that would explain the difference between the treatment arms, may then be calculated. This method cannot find the potential unmeasured confounder, but can provide insight into whether such a confounder is likely.

EXAMPLES OF OBSERVATIONAL STUDIES

An observational study is likely to be found to be credible where investigators, clinicians, and other stakeholders feel that treatment selection bias is likely to be minimized. An example is closure devices for vascular

procedures.[12] Approval studies for vascular closure devices have generally been small studies, often randomized trials, for evaluation of efficacy. Larger trials for evaluation of safety have not been mounted. To evaluate safety, the Food and Drug Administration sponsored a study with the American College of Cardiology (ACC) using the National Cardiovascular Data Registry (NCDR). The study population comprised 13,878 patients from 59 institutions. Multivariable analysis, controlling for demographic, physiologic, and procedural variables in addition to comorbidity, found that one device among several had high risk for any vascular complication compared with manual compression (odds ratio 2.38, $P = .0004$). This study was generally acknowledged as reasonable, and the problematic closure device was soon off the market. Another study with closure devices was received with greater skepticism. Thirty-day mortality was compared between closure devices and manual compression in a study of 271,845 patients undergoing percutaneous coronary intervention in the United Kingdom.[13] Unadjusted 30-day mortality was lower in the vascular closure group compared with manual pressure (hazard ratio [HR] 0.58; 95% CI, 0.54-0.61; 1.4% vs 2.4%, $P < .0001$), with reduced effect after propensity score correction (HR 0.91; 95% CI, 0.86-0.97; 1.8% vs 2.0%, $P < .001$). A source of confounding in this study is that the use of vascular closure devices is determined, among other things, by the suitability of the femoral artery based on the extent of peripheral vascular atheroma or calcification—variables that may also be associated with worse outcomes. In contrast, Wimmer et al[14] used an instrumental variable, in this case whether percutaneous coronary intervention operators were frequent or infrequent users of closer devices. Using data from NCDR, 2,056,585 percutaneous coronary interventions were performed by 4331 operators. Acute closure devices were used in 5.9% of low-use operators and 90.3% of high-use operators. There was little difference in clinical covariates. Unlike the results observed in the UK study, there was no difference in mortality with closure devices; absolute risk reduction 0.03% (95% CI, –0.07-0.04). There was a difference in access site bleeding; absolute risk reduction 0.36% (95% CI, 0.31-0.42). Of particular interest, the study employed nonaccess site bleeding as a "falsification endpoint," with an absolute risk reduction of 0.04% (95% CI, 0.01-0.07).[15] The falsification variable is important because if a clinically significant reduction in nonaccess site bleeding (such as gastrointestinal bleeding) had been observed despite no possible biological relationship with closure device use, then the reduction in access site bleeding would likely be the result of residual confounding and not due to the closure device. For evaluation of a treatment whose selection is heavily governed by variables that are of prognostic significance, instrumental variable-based methods may be superior to those using regression or propensity score methods.

Another example of problems with observational data can be seen in the comparison of drug-eluting and non-drug-eluting stents for treatment of coronary artery disease. Using NCDR, Douglas et al[16] linked outcomes of 262,700 (217,675 drug-eluting stent, 45,025 bare metal stent) patients undergoing percutaneous coronary intervention to Medicare claims data. Using propensity score analysis and inverse probability weighting, at 30 months, drug-eluting stent patients had a lower incidence of death (13.5% vs 16.5%; HR 0.75; 95% CI, 0.72-0.79; $P < .001$), with a smaller difference in additional revascularization (HR 0.91; 95% CI, 0.87-0.96). These results conflicted with previously published randomized trial data, which had not shown mortality benefits with drug-eluting stents. Kirtane et al[17] contrasted clinical trial and observational data drug-eluting stents and bare metal stents in a pair of meta-analyses, with data from 9,470 patients in 22 randomized trials and 182,901 patients in 34 observational studies. In the randomized trials, there were no differences in overall mortality between drug-eluting stents and bare metal stents (HR 0.97; 95% CI, 0.81-1.15; $P = .72$), but a significant reduction in target vessel revascularization (HR 0.45; 95% CI, 0.37-0.54; $P < .0001$). In observational studies, however, there was a significant reduction in mortality (HR 0.78; 95% CI, 0.71-0.86), in addition to the expected reduction in target vessel revascularization with drug-eluting stents (HR 0.54; 95% CI, 0.48-0.61). In the case of mortality, the results strongly suggested residual treatment selection bias in the observational studies. The data on additional revascularization is somewhat inconsistent within the observational studies, but clear in the randomized trials and consistent with the mechanism of action of drug-eluting stents. These studies bring out a critical point about observational studies: size does not overcome bias.

Observational studies can also be consistent with randomized trials. An example of this is the ACC Foundation and Society of Thoracic Surgeons Database Collaboration on the Comparative Effectiveness of Revascularization Strategies (ASCERT) study comparing coronary artery bypass grafting (CABG) and percutaneous coronary intervention in chronic stable ischemic heart disease. Weintraub et al[18] linked the ACC NCDR and the Society of Thoracic Surgeons Adult Cardiac Surgery Database to claims data from the Centers for Medicare and Medicaid Services for the years 2004 through 2008. Outcomes were compared using propensity scores and inverse probability weighting. Among patients ages >65 years with 2- or 3-vessel coronary artery disease, 86,244 underwent CABG and 103,549 underwent percutaneous coronary intervention. At 1 year, there was no significant difference in adjusted mortality between the groups (6.24% in the CABG group as compared with 6.55% in the percutaneous coronary intervention group; relative risk [RR] 0.95; 95% CI, 0.90-1.00). At 4 years, there was lower mortality with CABG than with percutaneous coronary intervention (16.4% vs 20.8%; RR 0.79; 95% CI,

0.76-0.82). The ASCERT study also contains an example of sensitivity analysis using the method of Lin et al.[11] An example was offered: if a confounder was present in 10% of the CABG and 35% of the percutaneous coronary intervention patients, and if it had a HR ≥2.09, then it could account for the observed difference in mortality between the study groups. Frailty could be that unmeasured confounder. However, the ASCERT results are largely consistent with clinical trial data. In a meta-analysis of 6 randomized trials with 6065 patients, Sipahi et al[15] found a long-term mortality benefit of CABG compared with percutaneous coronary intervention (RR 0.73; 95% CI, 0.62-0.86, $P < .001$). When randomized clinical trials and observational studies find similar results, it can help resolve concerns about the generalizability of randomized trial results to a broader population.

CAUSALITY

The fundamental consideration of evaluation of therapy is causality, that is, we want to know whether a therapy *causes* an observed improvement in outcome. While randomized trials have their limitations, they remain the gold standard for assessment of causality because it remains the best method for overcoming treatment selection bias. Observational studies have proven more difficult. In the 1950s, Doll and Hill published a pair of seminal papers showing an association between cigarette smoking and the development of lung cancer.[19,20] These papers, plus a study from Framingham on the impact of cigarette smoking on cardiovascular disease risk,[21] led to the first US Surgeon General's Report on Smoking and Health.[22,23] Nonetheless, controversy persisted over whether cigarette smoking was causally related to lung cancer as opposed to just being an association. In response, Hill developed a set of criteria to assess causal relationship between an exposure and future events.[24] These criteria can be applied to therapy that may reduce bad outcomes, as well as to an environmental exposure that increases bad outcomes. The Bradford Hill criteria are 1) temporal relationship—the cause must always come before the effect; 2) strength of association; 3) dose–response relationship; 4) consistency of the relationship; 5) biological plausibility; 6) consideration of alternatives; 7) experimental verification; 8) specificity (a specific cause for a specific effect); and 9) coherence (compatibility with existing knowledge).[24,25] Specificity is often omitted because it is generally not fulfilled for diseases that may have multiple causes. Establishing causality requires consideration of these criteria and then general acceptance by the scientific community, subject matter experts, and society at large. For risk factors, such as cigarette smoking, there is generally no choice; epidemiologic principles to assess causality must be considered. For great therapy, for Isoniazid for tuberculosis prior to antibiotic resistance, observational data alone could meet the Bradford Hill criteria and be adopted by the community. For more marginal therapies, the type that physicians consider every day, such as percutaneous coronary intervention for acute coronary syndromes, it is not clear that observational data alone can offer a convincing alternative to randomized trials that will fulfill the Bradford Hill criteria, pass regulatory hurdles, and be generally accepted by societal stakeholders.

CONCLUSIONS

Observational studies and randomized trials will continue to be used to assess therapies and diagnostic strategies. Results are most compelling where they find complementary results. It is difficult for observational studies to offer convincing evidence of causality, which limits their use for regulatory purposes. This limitation is primarily due to potential residual confounding of uncertain severity. Statements in observational studies such as "while residual selection bias cannot be excluded, we controlled for all measured differences between the treatment groups," should not be viewed as reassurance that the results can be interpreted to show a causal relationship between the intervention and outcome. Proper integration of results from all types of studies requires careful methods, modest conclusions, and the ability to consider new information.

References

1. Glesby MJ, Hoover DR. Survivor treatment selection bias in observational studies: examples from the AIDS literature. *Ann Intern Med*. 1996;124:999-1005.
2. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. 2002;288:321-333.
3. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med*. 1991;324:781-788.
4. D'Agostino RB Jr. Propensity scores in cardiovascular research. *Circulation*. 2007;115:2340-2343.
5. Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Med Care*. 2007;45:S103-S107.
6. Elze MC, Gregson J, Baber U, et al. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *J Am Coll Cardiol*. 2017;69:345-357.
7. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. 2007;297:278-285.
8. Burgess S, Timpson NJ, Ebrahim S, Davey Smith G. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol*. 2015;44:379-388.
9. Strong A, Rader DJ. Clinical implications of lipid genetics for cardiovascular disease. *Curr Cardiovasc Risk Rep*. 2010;4:461-468.

10. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc*. 1995;90:443-450.
11. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;54:948-963.
12. Tavris DR, Dey S, Albrecht-Gallauresi B, et al. Risk of local adverse events following cardiac catheterization by hemostasis device use—phase II. *J Invasive Cardiol*. 2005;17:644-650.
13. Farooq V, Goedhart D, Ludman P, et al. Relationship between femoral vascular closure devices and short-term mortality from 271 845 percutaneous coronary intervention procedures performed in the United Kingdom between 2006 and 2011: a propensity score-corrected analysis from the british cardiovascular intervention society. *Circ Cardiovasc Interv*. 2016;9(6). doi:10.1161/CIRCINTERVENTIONS.116.003560
14. Wimmer NJ, Secemsky EA, Mauri L, et al. Effectiveness of arterial closure devices for preventing complications with percutaneous coronary intervention: an instrumental variable analysis. *Circ Cardiovasc Interv*. 2016;9(4):e003464.
15. Sipahi I, Akay MH, Dagdelen S, Blitz A, Alhan C. Coronary artery bypass grafting vs percutaneous coronary intervention and long-term mortality and morbidity in multivessel disease: meta-analysis of randomized clinical trials of the arterial grafting and stenting era. *JAMA Intern Med*. 2014;174:223-230.
16. Douglas PS, Brennan JM, Anstrom KJ, et al. Clinical effectiveness of coronary stents in elderly persons: results from 262,700 Medicare patients in the American College of Cardiology-National Cardiovascular Data Registry. *J Am Coll Cardiol*. 2009;53:1629-1641.
17. Kirtane AJ, Gupta A, Iyengar S, et al. Safety and efficacy of drug-eluting and bare metal stents: comprehensive meta-analysis of randomized trials and observational studies. *Circulation*. 2009;119:3198-3206.
18. Weintraub WS, Grau-Sepulveda MV, Weiss JM, et al. Comparative effectiveness of revascularization strategies. *N Engl J Med*. 2012;366:1467-1476.
19. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J*. 1950;2:739-748.
20. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J*. 1954;1:1451-1455.
21. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd. Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study. *Ann Intern Med*. 1961;55:33-50.
22. Smoking and Health. U.S. Department of Health Education and Welfare. U.S. Government Printing Office; 1964.
23. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health, US Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Rockville, MD: US Department of Health and Human Services, Public Health Service, Office of the Surgeon General; 2014.
24. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295-300.
25. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. New York: Van Nostrand Reinhold Company; 1982.