Randomized Trials Analyzed as Observational Studies

Miguel A. Hernán, MD; Sonia Hernández-Díaz, MD; and James M. Robins, MD

Despite what you may have heard, randomized trials are not always free of confounding and selection bias. Randomized trials are expected to be free only from baseline confounding but not from postrandomization confounding and selection bias (1). In this commentary, we describe the settings in which postrandomization confounding and selection bias emerge in randomized trials, discuss the shortcomings of intention-to-treat analyses to handle these biases, and direct readers to more appropriate methods.

The neglect of postrandomization confounding and selection bias in randomized trials is the historical consequence of the fact that many early trials were short, small, double-blinded, tightly controlled experiments in highly selected patients. Most premarket trials still fit this description. In these experiments, randomization makes baseline confounding unlikely, whereas double-blinding, tight control, and short duration minimize postrandomization confounding (due to deviations from protocol or differential use of concomitant therapies) and selection bias (due to differential loss to follow-up). Such trials may be optimal to detect small treatment benefits but not to guide clinical decision making: follow-up too short for clinically relevant outcomes, patients unrepresentative, interventions unrealistic, sample size too small to identify adverse events.

A different breed of randomized trial is increasingly used to study the long-term effects of sustained clinical interventions in typical patients and care settings. These trials are more vulnerable to postrandomization confounding and selection bias. As an example, suppose we want to estimate the effect of estrogen plus progestin hormone therapy on the 5-year risk for breast cancer among postmenopausal women. We might consider an open-label randomized trial in which thousands of women within 5 years of menopause, with no history of cancer and no prior hormone therapy use, are randomly assigned to hormone therapy or no therapy. During follow-up, some women are observed to discontinue or start hormone therapy or concomitant therapies. They may also become lost to follow-up.

In this type of trial, sometimes referred to as a pragmatic or large, simple trial (2), confounding may arise from nonadherence if postrandomization prognostic factors (other than toxicity or contraindications) that affect treatment decisions are unequally distributed across groups, and selection bias may occur from loss to follow-up if prognostic factors affect decisions to stay in the study. That is, randomized trials of sustained interventions over long periods are subject to biases that we have learned to associate exclusively with observational studies.

The description of this pragmatic trial could also fit an observational study. We only need to replace "are randomly assigned to" with "decide to take." Apart from baseline randomization, there may be no differences between observational studies and randomized trials. Indeed, large, simple trials are designed to closely resemble observational studies. (Of course, observational studies, unlike large randomized trials, require adjustment for baseline confounders.)

Notwithstanding their similarities, the primary analysis of most randomized trials is "intention to treat," whereas that of many observational studies is "as treated." Why? A common justification is that an intention-to-treat analysis does not require adjustment for postrandomization factors because it estimates the effect of assigned (baseline) treatment. Although almost correct (adjustment for selection bias due to differential loss to follow-up is still required for validity), this argument begs the question of whether the intention-to-treat analysis estimates the effect of interest.

The answer is clearly "no" for safety trials. Take the Women's Health Initiative, a double-blind, randomized trial of estrogen plus progestin (3). The intention-to-treat hazard ratio of breast cancer was 1.25 (95% CI, 1.01 to 1.54) for hormone therapy versus placebo (3). An observational-type analysis of the trial based on inverse probability weighting estimated that the hazard ratio would have been 1.68 (CI, 1.24 to 2.28) if all women had followed the study protocol (4). Should a woman contemplating hormone therapy consider herself adequately informed if told that her risk for breast cancer will increase by 25%, when regular use may increase that risk by 68%? Worse, if the trial had included fewer women, the 95% CI from the intention-to-treat analysis would probably have included 1.0, which many would have incorrectly interpreted as lack of evidence of harm. Randomized clinical trials of safety outcomes that only report intention-to-treat estimates might be renamed as "randomized cynical trials" (5).

The answer is also "no" for many efficacy trials. Take the ACTG 002 (AIDS Clinical Trials Group 002), an early randomized trial in HIV-infected patients that compared high- versus low-dose zidovudine. The administration of prophylaxis therapy for *Pneumocystis* pneumonia, an opportunistic infection, was left to the physician's discretion. The intention-to-treat analysis suggested a survival benefit of low-dose zidovudine; however, individuals in the low-dose group received significantly more prophylaxis therapy than those in the high-dose group (61% vs. 50%). By the

This article was published online first at www.annals.org on 10 September 2013.

time the trial ended, prophylaxis for *Pneumocystis* pneumonia had become the standard of care. At that point, the relevant clinical question was whether the low-dose group would still have had better survival than the high-dose group had all trial participants received prophylaxis. This question is not addressed by an intention-to-treat analysis. An observational-type analysis of the trial based on g-estimation estimated a close to null survival benefit had all trial participants received prophylaxis (5).

In trials designed to estimate treatment benefits, a popular argument in support of the intention-to-treat analysis is that it estimates the efficacy (the effect of treatment under ideal conditions) in tightly controlled experiments and the effectiveness (the effect of treatment under realistic conditions) in pragmatic trials. However, a sharp distinction between efficacy and effectiveness is artificial and difficult to operationalize (6). After all, we do not try to distinguish between safety and "safetiness" in safety trials. Effectiveness, like safety, is a continuum that varies with the degree of adherence and other factors.

An alternative to the efficacy–effectiveness dichotomy is to be explicit about the effect of interest. For example, in the Women's Health Initiative hormone therapy trial, we might be interested in the per-protocol effect, which is the effect that would have been observed if the only deviations from the assigned treatment were for medical reasons specified in the protocol (such as toxicity or contraindications). In the ACTG 002 zidovudine trial, we might be interested in the controlled direct effect of low-dose zidovudine, which is the effect that would have been observed if all individuals had received prophylaxis for *Pneumocystis* pneumonia. Unfortunately, estimating the per-protocol and direct effects requires untestable conditions and, even when these conditions are true, the commonly used "per-protocol" and "as-treated" analyses may not provide valid estimates because they fail to appropriately account for postrandomization biases.

The good news is that there are methods, often referred to as g-methods, that appropriately adjust for postrandomization biases (2). These methods, developed by Robins and collaborators since 1986, require data on postrandomization treatment and covariates. A first group of g-methods—inverse probability weighting, g-estimation, and the parametric g-formula—provides valid per-protocol estimates under the same untestable assumptions that we usually reserve for observational studies (that is, all postrandomization prognostic factors that affect either treatment choices or loss to follow-up are correctly measured and modeled) (7). A second type of g-method—a form of g-estimation that generalizes instrumental variable estimation—does not require the same assumptions as observational studies but rather requires detailed modeling assumptions about the effect of treatment. If there is truly no treatment effect, there will be no difference between testing the null using this form of g-estimation or using an intention-to-treat analysis. The **Table** summarizes the conditions required for the validity of g-methods in randomized trials.

In summary, the similarities between follow-up studies with and without baseline randomization are becoming increasingly apparent as more randomized trials study the effects of sustained interventions over long periods in real-world settings. What started as a randomized trial may effectively become an observational study that requires analyses to complement and go beyond intention-to-treat analyses. A key obstacle in adoption of these complementary methods is a widespread reluctance to accept that overcoming the limitations of intention-to-treat analyses necessitates untestable assumptions. Embracing these more

Table. **Correctly Specified Models Required to Validly Estimate the Intention-to-Treat Effect and Effects Defined by Postrandomization Interventions, Including Per-Protocol and Direct Effects***

Model Type	Intention-to-Treat Effect	Effects Defined by Postrandomization Interventions			
	IP Weighting for Selection Bias	IP Weighting for Confounding and Selection Bias	g-Estimation for Confounding, IP Weighting for Selection Bias	Parametric g-Formula for Confounding and Selection Bias	g-Estimation (Instrumental Variables) for Confounding, IP Weighting for Selection Bias
Model for loss to follow-up given joint determinants of loss to follow-up and the outcome	Yes	Yes	Yes	No	Yes
Model for treatments given past treatments and confounders	No	Yes	Yes	No†	No
Structural (dose–response) model for outcome given the treatments of interest	No	Optional	Yes	No, but outcome model given treatment and confounders is required	Yes
Model for confounders given past treatments and confounders	No	No	No	Yes	No

IP = inverse probability.
* Modified from reference 8.
† Necessary to estimate some effects.

sophisticated analyses will require a new framework for both the design and conduct of randomized trials.

From Harvard School of Public Health, Boston, Massachusetts.

Grant Support: In part by grants R01 HL080644 and R01 AI102634 from the National Institutes of Health.

Potential Conflicts of Interest: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M13-1455.

Requests for Single Reprints: Miguel A. Hernán, MD, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115; e-mail, miguel_hernan@post.harvard.edu.

Current author addresses and author contributions are available at www.annals.org.

Ann Intern Med. 2013;159:560-562.

References

1. **Robins JM, Hernán MA.** Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, eds. Advances in Longitudinal Data Analysis. New York: Chapman & Hall/CRC Pr; 2009.

2. **Lesko SM, Mitchell AA.** The use of randomized controlled trials for pharmacoepidemiologic studies. In: Strom BL, Kimmel SE, Hennessy S, eds. Pharmacoepidemiology. West Sussex, UK: Wiley-Blackwell; 2012: 553-99.

3. **Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al; Writing Group for the Women's Health Initiative Investigators.** Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321-33. [PMID: 12117397]

4. **Toh S, Hernández-Díaz S, Logan R, Robins JM, Hernán MA.** Estimating absolute risks in the presence of nonadherence: an application to a follow-up study with baseline randomization. Epidemiology. 2010;21:528-39. [PMID: 20526200]

5. **Robins JM, Greenland S.** Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc. 1994;89:737-49.

6. **Hernán MA, Hernández-Díaz S.** Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9:48-55. [PMID: 21948059]

7. **Robins JM.** Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, eds. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag; 1999:95-134.

8. **Toh S, Hernán MA.** Causal inference from longitudinal studies with baseline randomization. Int J Biostat. 2008;4:Article 22. [PMID: 20231914]

Current Author Addresses: Drs. Hernán, Hernández-Diaz, and Robins: Harvard School of Public Health, Department of Epidemiology, 677 Huntington Avenue, Boston, MA 02115.

Author Contributions: Conception and design: M.A. Hernán, S. Hernández-Díaz, J.M. Robins.
Drafting of the article: M.A. Hernán, S. Hernández-Díaz, J.M. Robins.
Critical revision for important intellectual content: M.A. Hernán, S. Hernández-Diaz, J.M. Robins.
Final approval of the article: M.A. Hernán, S. Hernández-Diaz, J.M. Robins.
Statistical expertise: M.A. Hernán, S. Hernández-Díaz, J.M. Robins.
Obtaining of funding: M.A. Hernán.
Administrative, technical, or logistic support: M.A. Hernán.