Published in final edited form as:
Ann Intern Med. 2013 October 15; 159(8): . doi:10.7326/0003-4819-159-8-201310150-00709.

Randomized trials analyzed as observational studies

Miguel A. Hernán[1,2,3], **Sonia Hernández-Díaz**[1], and **James M. Robins**[1,2]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, MA

[2]Department of Biostatistics, Harvard School of Public Health, Boston, MA

[3]Harvard-MIT Division of Health Sciences and Technology, Boston, MA

Despite what you may have heard, randomized trials are not always free of confounding and selection bias. Randomized trials are only expected to be free from baseline confounding, but not from post-randomization confounding and selection bias.[1] In this commentary we describe the settings in which post-randomization 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 post-randomization 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 pre-market trials still fit this description. In these experiments, randomization makes baseline confounding unlikely whereas double-blinding, tight control, and short duration minimize post-randomization confounding (e.g., due to deviations from protocol or differential use of concomitant therapies) and selection bias (e.g., 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 post-randomization 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 of breast cancer among postmenopausal women. We might consider an open label randomized trial in which thousands of women within five years of menopause, with no history of cancer and no prior hormone therapy use, are randomly assigned to hormone therapy or no therapy. During the 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 as a pragmatic or large simple trial[2]—confounding may arise from non-adherence if post-randomization prognostic factors (other than toxicity or contraindications) that affect treatment decisions are unequally distributed across arms,

Correspondence: Miguel Hernán. Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. Tel: 617 432 0101, miguel_hernan@post.harvard.edu.

and selection bias 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” by “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 post-randomization factors because it estimates the effect of assigned (baseline) treatment. While 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 double-blind randomized trial of estrogen plus progestin. The intention-to-treat hazard ratio (95% CI) of breast cancer was 1.25 (1.01, 1.54) for hormone therapy versus placebo.[3] An observational-type analysis (inverse probability weighting, see below) of the trial estimated that the hazard ratio would have been 1.68 (1.24, 2.28) if all women had followed the study protocol.[4] As a woman considering regular use of hormone therapy, would you consider yourself adequately informed if told that your breast cancer risk will increase by 25% when regular use may increase risk by 68%? Worse, if the trial had included fewer women, the 95% CI from the intention-to-treat analysis would have likely included 1, 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 an early randomized trial in HIV-infected patients, the ACTG 70, which compared high- versus low-dose zidovudine. The administration of prophylaxis therapy for PCP, an opportunistic infection, was left to the physicians’ discretion. The intention-to-treat analysis suggested a survival benefit of low-dose zidovudine. However, individuals in the low-dose arm received significantly more prophylaxis therapy than those in the high-dose arm (61% versus 50%). By the time the trial ended, prophylaxis for PCP had become the standard of care. At that point, the relevant clinical question was whether the low-dose arm would still have had better survival than the high-dose arm had all trial participants received prophylaxis. This question is not addressed by an intention-to-treat analysis. An observational-type analysis (g-estimation, see below) of the trial estimated a close to null survival benefit had all trial participants received prophylaxis.[5]

In trials that 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, in safety trials we do not try to distinguish between safety and “safetiness”. Effectiveness, like safety, is a continuum that varies with 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 WHI hormone therapy trial we might be interested in the per-protocol effect, that is, the effect that would have been observed if the only deviations from the assigned treatment were for medical reasons specified in the protocol (e.g., toxicity, contraindications), and in the ACTG 70 zidovudine trial we might be interested in the controlled direct effect of low-dose zidovudine, that is, the effect that would have been observed if all individuals had received prophylaxis for PCP. Unfortunately, estimating 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 post-randomization biases.

The good news is that there exist methods that appropriately adjust for post-randomization biases.[2] These so-called g-methods, developed by Robins and collaborators since 1986, require data on post-randomization 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, i.e., all post-randomization 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 second type of g-method 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 that complement, but go beyond, intention-to-treat analyses. A key obstacle in the 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 sophisticated analyses will require a new framework for both the design and conduct of randomized trials.

Acknowledgments

This work was partly funded by NIH grants R01 HL080644 and R01 AI102634.

References

1. Robins, JM.; Hernán, MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. Advances in Longitudinal Data Analysis. New York: Chapman and Hall/CRC Press; 2009.
2. Lesko, SM.; Mitchell, AA. The use of randomized controlled trials for pharmacoepidemiologic studies. In: Strom, BL.; Kimmel, SE.; Hennessy, S., editors. Pharmacoepidemiology. West Sussex, England: Wiley-Blackwell; 2012.
3. Writing Group for the Women’s Health Initiative Investigators. Risks and benefits of estrogen plus progestin therapy in healthy postmenopausal women: principal results from the Women’s Health Initiative randomized controlled trial. JAMA. 2002; 288:321–33. [PubMed: 12117397]
4. Toh S, Hernández-Díaz S, Logan RW, et al. Estimating absolute risks in the presence of non-adherence: an application to a follow-up study with baseline randomization. Epidemiology. 2010; 21:528–39. [PubMed: 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. Journal of the American Statistical Association. 1994; 89:737–749.

6. Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. Clinical Trials. 2012; 9(1):48–55. [PubMed: 21948059]
7. Robins, JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, ME.; Berry, D., editors. Statistical Models in Epidemiology: The Environment and Clinical Trial. New York: Springer-Verlag; 1999. p. 95-134.

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*

	Intention-to treat effect	Per-protocol effects, direct effects, and other effects defined by post-randomization interventions			
	Intention-to-treat analysis with IP weighting to correct 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 a treatment given past treatment 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
Models for confounders given past treatments and confounders	No	No	No	Yes	No

*Modified from Toh S, Hernán MA. Causal inference from longitudinal studies with baseline randomization. *International Journal of Biostatistics* 2008; 4(1):1–30. Article 22. Available at: http://www.bepress.com/ijb/vol4/iss1/22.

†Necessary to estimate some effects