European Journal of Epidemiology (2019) 34:715–718
https://doi.org/10.1007/s10654-019-00516-3

ESSAY

Generalizing from the results of randomized studies of treatment: Can non-randomized studies be of help?

Noel S. Weiss[1]

Received: 16 August 2017 / Accepted: 8 September 2017 / Published online: 11 April 2019
© Springer Nature B.V. 2019

Different patients who have the same illness may receive different forms of treatment. By documenting their treatment and the progression and complications of their illness, an attempt can be made to draw inferences regarding the relative impact of these different treatments. However, the interpretation of such comparisons can be compromised by underlying differences in the likelihood of the various outcome events among the patient treatment groups. Differences of opinion exist with respect to the frequency and magnitude of the confounding that may arise from these underlying differences. Benson and Hartz [1] concluded that "misuse of observational studies [of the efficacy of therapy] does not often occur in the recent literature", and that non-randomized studies of therapy "usually do provide valid information." On the other hand, Pocock and Elbourne [2] state that greater reliance on the results of nonrandomized studies of therapeutic efficacy could lead to "considerable dangers to clinical research and even to the well-being of patients." Particularly strident on this point was an editorial [3] which opined that "In an era of evidence-based medicine, controlled clinical trials are required to introduce new therapies into the clinic."

Previously, I have argued for a middle ground [4], suggesting that the results of nonrandomized studies of treatment effects are likely to provide valid information when: (a) there is a large observed difference in the incidence of the health outcome in question between treatment groups; and (b) the issue of confounding can be dealt with adequately (in the study design and/or analysis). Because people differ in their perception of what is "large" and what is "adequate", it is not surprising that there can be differences of opinion regarding the interpretation of a given non-randomized study of treatment effects.

The issue of the validity of non-randomized studies of treatment has been brought into focus by concerns regarding the generalizability of the results of randomized trials of therapy. For example, in the area of treatment of cancer, Elting [5] has contended that randomized trials "are not sufficient to improve therapy among subpopulations that are excluded from trials", and advocates the conduct of "observational studies of the effectiveness of cancer therapies across the entire population to extend the benefits of new therapies to all cancer patients."

In this commentary, I will describe circumstances in which results obtained in randomized trials of treatment are and are not more broadly applicable, and the potential pitfalls of a too-casual reliance on non-randomized studies to gauge therapeutic impact in the types of persons who were not represented in the trials' study populations.

There can be limitations to the generalizability of the results of randomized trials, due to…

1. Differences between the trial participants and the broader population regarding characteristics that influence the impact of the therapy under consideration

Situations exist in which there is reason to believe that the findings of a randomized trial will not be applicable to a different population, or to a subgroup of persons not well represented in the trial [6]. For example, among patients with congestive heart failure who were receiving an angiotensin-converting enzyme (ACE) inhibitor, randomization to a potassium-sparing diuretic was associated with a decrease in all-cause mortality without an appreciable increase in the incidence of hyperkalemia [7]. However, patients with a limited ability to regulate electrolyte levels—those with impaired renal function—were excluded from the trial, and so there would be a basis for concern that the findings of the trial with respect to the incidence of hyperkalemia might not

✉ Noel S. Weiss
nweiss@uw.edu

[1] University of Washington, Seattle, USA

be applicable to a broader range of patients. In a later, non-randomized, study of hyperkalemia among Ontario residents ages 65 years and older who had been receiving an ACE inhibitor (which included those with and without evidence of renal disease) [8], a higher proportion of cases (8.2%) than controls (0.3%) had been prescribed a potassium-sparing diuretic during the prior week (adjusted odds ratio = 20.3, 95% confidence interval = 13.4–30.7). That the incidence of hyperkalemia truly was elevated among elderly persons in Ontario with heart failure who took both an ACE inhibitor and a potassium-sparing diuretic is suggested not only by the strong association seen, but also by the investigators' ability to adjust for the presence of renal failure and for the use of other medications that bear on potassium metabolism.

Differences between trial participants and other persons are not themselves evidence that the trial results do not apply more broadly. For example, a double-blind randomized trial among men who have sex with men documented a sharp reduction in the acquisition of HIV infection associated with receipt of pre-exposure prophylaxis (PrEP) with anti-retroviral drugs [9]. However, there was concern that the results might not generalize well outside the trial population, in which men receiving PrEP would know their treatment status and thus might adopt high-risk sexual practices that could compromise the benefit that the drugs provide. This concern was addressed in a subsequent randomized study [10] in which there was no blinding as to treatment assignment, so as to reflect the "real world" experience. The estimate of PrEP efficacy from that study—similar to the estimate from the blinded trial—is evidence that the hypothesized behavioral disinhibition associated with knowledge of PrEP status did not materially bear on the more general applicability of the results of the original trial.

Among patients being treated at a large cancer center [11], those entered into randomized trials had (on average) different demographic characteristics and features of their malignancies than did patients who were not entered into those trials. While the authors of that study suggested that this "calls into question the generalizability of trial results" obtained at that center, treatment efficacy may or may not differ according to these characteristics and features. Concern regarding generalizability should be prompted by specific reasons for suspecting that the benefit (or lack thereof) observed in the trial ought to be dissimilar between patients with and without a given characteristic.

Similarly, Dutch investigators [12] had concerns regarding the generalizability of the results of their randomized trial of different approaches to adjuvant therapy for breast cancer. Among women 75 years and older in the trial, all-cause mortality was 28% lower than among other Dutch women with breast cancer who were comparable to the trial participants regarding both demographic and tumor characteristics. No similar mortality difference was seen for trial participants ages 65–74 years. The authors concluded that the "trial participants aged 75 years and above do not represent elderly breast cancer patients of corresponding age from the general population, which hampers the external validity" of their trial. In my opinion, the all-cause mortality difference has little bearing on the issue of external validity. Unless there are specific reasons to suspect that, among women over 75 years of age, the size of the survival benefit associated with a particular form of adjuvant therapy differs depending on the underlying likelihood of survival, I would believe that the evidence of benefit obtained in the trial would be a reflection of that present among elderly Dutch women in general.

2. A difference in the nature of the intervention (or the means by which it is administered) between the trial and community setting

In the trial of potassium-sparing diuretics among patient with congestive heart failure described earlier [7], there was close monitoring of serum potassium levels. This safeguard enabled dose reduction or cessation of the diuretic in patients with rising potassium levels, before the development of serious hyperkalemia. Such close monitoring would be expected to take place relatively less often outside the trial setting, and so is another possible explanation for the disparity between the results of the randomized trial and the Ontario case–control study [8] regarding the incidence of hyperkalemia.

A pooled analysis of five randomized trials of warfarin anticoagulation in patients with chronic atrial fibrillation observed an annual incidence of major hemorrhage of 1.3 per 100 among patients in the active treatment arm of the trial, a 30% increase above the rate among patients in the placebo arm [13]. Commentators on this analysis suggested, with good reason, that the experience of the warfarin-treated patients in the trials "may underestimate the true risk of hemorrhage in clinical practice [since] the mechanisms used in the trials [to measure the level of anti-coagulation] may be more reliable than those used in routine practice" [14].

For patients not well-represented by participants in randomized trials, can we obtain guidance from the results of non-randomized studies?

Some have suggested that non-randomized studies of the efficacy or safety of a particular therapy can complement randomized trials of that same therapy. For example, recently a recommendation has been made for the creation of "an archive of patient profiles using data from all study types and data sources [from which] the clinician seeking guidance for the management of an individual patient will… find approximate matches in the archive that describe how similar patients responded to a contemplated treatment and alternative treatments" [15]. Others have asserted [16] that though "clinical trials must remain the gold standard for identifying effective strategies for promoting health and managing disease, … large-scale observational studies have value in … confirming results in understudied patient subsets." This sentiment has been elaborated upon [17]: "Clinical trials select only a small, artificial subset of the population. A regular, ordinary person who walks into a doctor's office doesn't usually fit…only occasionally have you got a clinical-trial-based guideline facing you right now. [Fortunately] tons of applicable evidence are locked away in health systems' electronic medical records." However, the "tons of applicable evidence" that can be derived from "large-scale observational studies" often will not provide a valid estimate of the consequences of a given therapy in a given patient or subgroup of patients. As has been noted previously [18], in many instances the large-scale data bases that are increasingly available to evaluate potential treatment impact are limited with regard to the information that is present regarding the specifics of treatment and patient outcomes, and also with regard to potential confounding factors.

1. Specifics of treatment. The same databases in Ontario, Canada, that were so successful in identifying the increased risk of hyperkalemia in relation to receipt of combined use of an ACE inhibitor and a potassium-sparing diuretic were unable to contribute useful information on the impact of screening colonoscopy on risk of fatal colorectal cancer in that same population [19, 20]. In the available electronic data, though the receipt of colonoscopy could be ascertained accurately, there was no information regarding whether or not the test was done for purposes of screening or, instead, in response to symptoms or signs of colorectal cancer.
2. Specifics of illness outcomes. For some conditions (e.g., hip fracture) the records in some large data bases appear to have a high degree of accuracy [21], but for some others (e.g., venous thromboembolism) it is clear that they do not [22]. Because many data bases are generated from billing records, they may not contain information regarding the subclasses of a disease (e.g., tumor molecular subtype) that can be important in studies of etiology or prognosis.
3. Potential confounding variables. As noted at the outset of this essay, there are times when the specter of confounding can be dealt with successfully in non-randomized studies: Perhaps a characteristic associated with disease outcome is not materially related to receipt of a particular form of therapy, or it is related but the characteristic can be measured well and adjusted for in the data analysis. But experience has taught us to be careful here. For example, electronic data from a large pre-paid health care plan were used to examine the possibility of a reduced risk of all-cause mortality in relation to receipt of influenza vaccine during the prior year [23]. Despite being able to ascertain and adjust for the presence of various conditions that plausibly could be predictors both of mortality and receipt of vaccination (e.g., heart disease, lung disease), the study observed a strikingly low risk of death among vaccine recipients prior to the start of flu "season". It is virtually certain that influenza vaccination did not exert a true benefit during that period of time. More plausibly, the association was the result of "residual" confounding, due to an association of extreme frailty—which could not be assessed in the available data—with both mortality and the non-receipt of immunization that year.

Even when a database contains information on one or more confounding variables, it may not be for the point in time at which the treatment being evaluated has been initiated. For example, population-based cancer registries obtain detailed information regarding the characteristics of a patient's malignancy (such as stage and grade) at the time of diagnosis, but generally not afterwards. Therefore the results of a registry-based study of, say, the use of anti-depressant medications initiated months or years following cancer diagnosis in relation to survival would be difficult to interpret, due to the inability to adjust for tumor progression between the time of diagnosis and the time when medication had been started.

The interpretation of non-randomized studies of therapy conducted among types of persons not well-represented in earlier trials of such therapy can be enhanced by including, as a separate stratum, persons who were well-represented in those trials. For example, a meta-analysis of trials among patients with early-stage non-small cell lung cancer, most younger than 70 years, observed an 11% mortality reduction associated with randomization to receive adjuvant chemotherapy [24]. In a non-randomized study of the same therapy, conducted within the US Veterans Health Administration, Ganti et al. [25] included patients older and younger than 70 years. Receipt of adjuvant chemotherapy was associated with about a 20% mortality reduction among persons in both age groups. Though the results obtained in this latter study appear to overestimate the survival benefit of adjuvant chemotherapy for early-stage non-small cell lung cancer, they suggest that whatever benefit is present is similar in persons older and younger than 70 years.

References

1. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. N Engl J Med. 2000;342:1878–86.
2. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? N Engl J Med. 2000;342:1907–9.
3. Goldberg IJ. To drink or not to drink? N Engl J Med. 2003;348:163–4.
4. Weiss NS, Clinical epidemiology: the study of the occurrence of illness. Oxford, New York; 2006.
5. Elting LS. Author reply. Cancer. 2007;109:342.
6. Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. Arch Intern Med. 2008;168:133–5.
7. Pitt B, Zannad F, Remme WJ, et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. N Engl J Med. 1999;341:709–17.
8. Juurlink DN, Mamdani M, Kopp A, et al. Drug–drug interactions among elderly patients hospitalized for drug toxicity. JAMA. 2003;289:1652–8.
9. Grant RM, Lama JR, Anderson PI, et al. Pre-exposure prophylaxis for HIV prevention in men who have sex with men. N Engl J Med. 2010;363:2587–99.
10. McCormack S, Dunn DT, Sesai M, et al. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomized trial. Lancet. 2016;387:53–60.
11. Elting LS, Cooksley C, Bekele BN, et al. Generalizability of cancer clinical trial results: prognostic differences between participants and nonparticipants. Cancer. 2006;106:2452–8.
12. Van de Water W, Kiderlen M, Bastiaannet F, et al. External validity of a trial comprised of elderly patients with hormone receptor-positive breast cancer. JNCI. 2014;106:1–8.
13. Atrial Fibrillation Investigators. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation: analysis of pooled data from five randomized controlled trials. Arch Intern Med. 1994;154:1449–57.
14. Stern S, Altkorn D, Levinson W. Anticoagulation for chronic atrial fibrillation. JAMA. 2000;283:2901–3.
15. Horwitz RI, Singer BH. Why evidence-based medicine failed in patient care and medicine-based evidence will succeed. J Clin Epidemiol. 2017;84:14–7.
16. Lauer MS, Collins FS. Using science to improve the nation's health system: NIH's commitment to comparative effectiveness research. JAMA. 2010;303:2182–3.
17. Goldman B. Treatments that work for people just like you. Winter: Stanford Medicine; 2016. p. 32.
18. Weiss NS. The new world of data linkages in clinical epidemiology: Are we being brave or foolhardy? Epidemiology. 2011;22:292–4.
19. Baxter NN, Goldwasser MA, Paszat LF, et al. Association of colonoscopy and death from colorectal cancer. Ann Intern Med. 2009;150:1–8.
20. Weiss NS, Dhillon PK, Etzioni R. Case-control studies of the efficacy of cancer screening: overcoming bias from non-random patterns of screening. Epidemiology. 2004;15:409–13.
21. Ray WA, Griffin MR, Schaffner W, et al. Psychotropic drug use and risk of fracture. N Engl Med. 1987;316:363–9.
22. Gerstman BB, Friesman JP, Hine LK. Use of subsequent anticoagulants to increase the predictive value of Medicaid deep venous thromboembolism diagnosis. Epidemiology. 1990;1:122–7.
23. Jackson LA, Jackson ML, Nelson JC, et al. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. Int J Epidemiol. 2006;35:337–44.
24. Pignon JP, Tribodet H, Scagliotti GV, et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 2008;26:3552–9.
25. Ganti AK, Williams CD, Gajra A, et al. Effect of age on the efficacy of adjuvant chemotherapy for resected non-small cell lung cancer. Cancer. 2015;121:2578–85.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.