VALUE IN HEALTH 15 (2012) 215–216

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

EDITORIAL

Can Observational Studies Approximate RCTs?

"It is the position of this Task Force that rigorous well designed and well executed Observational Studies (OS) can provide evidence of causal relationships" [1]. All flows from this carefully crafted statement in the middle of the ISPOR Good Research Practices Task Force Report, which provides a well-reasoned and well-conceived summary of the potential role in comparative effectiveness research (CER) of observational studies, and especially of prospective studies—ones that collect information. But the conceptual basis for the value of OS and the principles articulated would lack compelling interest if fulfilling the goals of CER did not force us to make use of OS, and, in doing so, bring OS as close to parity as possible with randomized controlled trials (RCTs).

The implication of this task force review article is that the question of RCTs versus OS should be reframed: Can conclusions from OS studies, if optimized ("rigorous, well-designed and well-executed"), be used, at approximately the same level of the hierarchy of quality, hopefully in conjunction with RCTs, in systematic reviews to inform clinical practice guidelines and policymakers about the effectiveness of medical practices.

In theory, few would argue that RCTs can establish causal relationships where OS cannot, but in clinical trials research the well-known drawbacks and flaws of RCTs, especially in CER, may cause us to pause in rating them higher than well-done OS, as aspired to in the creation of this report. Of course, disagreement will continue as to whether causal inferences can be drawn from OS with sufficient confidence to support either clinical or policy decisions. The potential problems with OS for CER are certainly greater than for OS of comparative safety, because of the greater likelihood of important selection biases. Yet, OS for comparative safety are still often discounted; despite multiple observational studies showing that rofecoxib increased the risk of acute myocardial infarction, the major stimulus to change was an RCT [2]. Thus, we need guidance such as this report's to articulate standards around which we can try to achieve a consensus.

The key, therefore, lies in the phrase "rigorous well-designed and well-executed," and this article provides an outline of the major determinants of high quality in crafting OS. We will summarize them below, but it deserves noting that, in past research, when neither RCTs nor OS were necessarily optimized, comparison of OS to RCTs revealed that there is a great deal of correspondence between the two [3]. Several major reviews comparing OS and RCTs have concluded that the results are similar [3–5] in a surprising percentage of studies of the same topic. Even in the often-quoted Women's Health Institute, as noted in the ISPOR report, when observational hormone replacement treatment data were analyzed only for treatment initiators, the results were essentially the same as in the random trial data, which itself enrolled only previously untreated patients [1]. In this case, the emphasis is on the subgroup of long-term versus naive users, rather than the more general topic of heterogeneity of treatment effects in varying subsets of patients with the same medical conditions. If OS were well conceived and well performed on a relatively homogeneous sample of patients, however, the persuasiveness of the results might be very high, especially in light of those problems increasingly found in RCTs: crossover, attrition, nonadherence, and varying quality of the sites and providers (or varying fidelity to an intervention), all of which serve to diminish the internal validity of the RCT.

Generally, it is a given that even in large pragmatic trials, OS can include a wider range of patients and can address rare events harms and multiple outcomes better than do RCTs. But the most contentious issue remains the internal validity. To elevate the level of internal validity, the task force makes a series of recommendations designed to bring the internal validity of OS closer and closer to that of RCTs.

First, the strength of prevailing opinion, the need for large sample sizes (as for harms), the value of studying multiple outcomes, the difficulties (adherence, crossovers, switching treatments, and time-varying covariates) encountered if randomization is contemplated, the need for early and highly generalizable answers—all these factors should be weighed both in deciding to perform an OS and in its design and conduct. Second is the unambiguous specification of the research question and the population to which potentially casual inferences can be aimed. They emphasize the notion of testing a hypothesis, as done in RCTs. The design issues and strategies to enhance the causal inferences are intended to have the OS approximate an RCT.

Given the potential for achieving this level of approximation after application of the strategies proposed by the task force, the credibility of the results of an optimal OS must then be balanced by the well-known and increasing flaws and drawbacks of RCTs, including not only the issues mentioned earlier but also the increasing problem of better "usual care" (sometimes called secular trends) and the problem of the average effect not revealing the often vastly different results of different subgroups within the same diagnosis [6–8].

The authors sensibly recommend such strategies as increasing the number of comparators, of examining outcomes that are not directly affected by the intervention, of prospectively including all the potential cofounders, and on focusing on the highest priority comparisons where there are multiple treatments. They also address confounders and bias, first by recognizing the biases and second by recommending strategies to deal with them: understanding the practice patterns that induce bias; distinguishing inception from already treated cohorts; using statistical approaches to better balance groups (propensity scores, instrument variables). They wisely point out [1] that the "absence of treatment heterogeneity is a crucial assumption for virtually all analytical approaches," the kind of statement that is not often heard in the disputes over study designs. They also give tips on sample size determination with

heterogeneity of treatment effects in mind—no clear answers but increasing awareness of the problems.

Perhaps the guidance's most important contribution is its focus on the ability to improve the validity of OS by prospectively collecting data that would otherwise be unavailable. Examples include quality-of-life data, information about medical history, or measures such as pulmonary function testing that may be missing or performed with unacceptable variability.

Finally, they focus on execution, where the practical issues may lead to matching or stratification either on the basis of a propensity score or, as some have suggested, on the basis of a prespecified and prestudied clinical variable, hopefully reduced by a multivariate composite [6–9]. This task force report is full of "pearls" and tips for the wise. It is also balanced and goes out of its way not to polarize the contentious debate between RCTs and OS. The concrete tips and the balanced tone encourage the readers to be as thoughtful as possible about the choice of an OS, emphasizing the notion that OS can be viewed as RCTs without randomization and that RCTs are OS once the randomization is complete. That kind of thinking moves us to go beyond the rigid definitions and hierarchies of designs and think about inferences, about how systematic review groups will rate the research, how clinical practice guidelines groups will use the research, and how CER can flourish and accomplish its goals.

This is such a fine achievement that criticism seems petty, but several directions for the future may be advised. One is to define the "strength" of preferences (strength of opinions) such that the choice of design will be easier. Another is to more directly address the problems of subgroups, how to decide whether to exclude them for other researchers to study, or to include them by matching or stratification, another is to seek ways in which both RCTs and OS can be undertaken simultaneously or in parallel, using similar or overlapping variables, interventions, outcomes, and covariates, so that conclusions can be drawn as were done belatedly in the Women's Health Institute and were attempted in the recent study of patent foramen ovales [5]. Because collection of new data and some of the recommendations impose substantial costs, it will also be helpful to have guidance about when these features are truly necessary. Because CER resources are constrained, we need all the advice we can get about allocating them wisely.

Sheldon Greenfield, MD
Health Policy Research Institute, University of California, Irvine, Irvine, CA, USA.

Richard Platt, MD
Harvard Pilgrim Health Care Institute, Boston, MA, USA.

1098-3015/$36.00 – see front matter

Published by Elsevier Inc.
doi:10.1016/j.jval.2012.01.003

REFERENCES

[1] Berger ML, Dreyer N, Anderson F, et al. Prospective observational studies to assess comparative effectiveness: The ISPOR Good Research Practices Task Force Report. Value Health 2012;15;217–30.
[2] Bresalier RS, Sandler RS, Quan H, et al.; Adenomatous Polyp Prevention on Vioxx (APPROVe) Trial Investigators. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092–102.
[3] Concato J, Lawler EV, Lew RA, et al. Observational methods in comparative effectiveness research. Am J Med 2010;123(12, Suppl. 1): e16–23.
[4] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–86.
[5] Kitsios GD, Dahabreh IJ, Abu Dabrh AM, et al. Patent foramen ovale closure and medical treatments for secondary stroke prevention: a systematic review of observational and randomized evidence. Stroke 2012;43:422–31. Epub 2011 Dec 15.
[6] Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. Am J Med 2007;120(Suppl.):S3–9.
[7] Kent DM, Rothwell PM, Loannidis JPA, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
[8] Greenfield S, Billimek J, Pellegrini F, et al. Comorbidity affects the relationship between glycemic control and cardiovascular outcomes in diabetes: a cohort study. Ann Intern Med 2009;151:854–60.
[9] Kaplan SH, Billimek J, Sorkin D, et al. Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. Med Care 2010;48(Suppl.):S9–16.