Benchmarking Observational Analyses Against Randomized Trials: a Review of Studies Assessing Propensity Score Methods

Shaun P. Forbes, AM[1,2] *and Issa J. Dahabreh, MD ScD*[1,2,3]

[1]Center for Evidence Synthesis in Health, Brown University School of Public Health, Providence, USA; [2]Department of Health Services, Policy & Practice, Brown University School of Public Health, Providence, USA; [3]Department of Epidemiology, Brown University School of Public Health, Providence, USA.

BACKGROUND: Observational analysis methods can be refined by benchmarking against randomized trials. We reviewed studies systematically comparing observational analyses using propensity score methods against randomized trials to explore whether intervention or outcome characteristics predict agreement between designs.

METHODS: We searched PubMed (from January 1, 2000, to April 30, 2017), the AHRQ Scientific Resource Center Methods Library, reference lists, and bibliographies to identify systematic reviews that compared estimates from observational analyses using propensity scores against randomized trials across three or more clinical topics; reported extractable relative risk (RR) data; and were published in English. One reviewer extracted data from all eligible systematic reviews; a second reviewer verified the extracted data.

RESULTS: Six systematic reviews matching published observational studies to randomized trials, published between 2012 and 2016, met our inclusion criteria. The reviews reported on 127 comparisons overall, in cardiology (29 comparisons), surgery (49), critical care medicine and sepsis (46), nephrology (2), and oncology (1). Disagreements were large (relative RR < 0.7 or > 1.43) in 68 (54%) and statistically significant in 12 (9%) of the comparisons. The degree of agreement varied among reviews but was not strongly associated with intervention or outcome characteristics.

DISCUSSION: Disagreements between observational studies using propensity score methods and randomized trials can occur for many reasons and the available data cannot be used to discern the reasons behind specific disagreements. Better benchmarking of observational analyses using propensity scores (and other causal inference methods) is possible using observational studies that explicitly attempt to emulate target trials.

Prior Presentations *Earlier versions of this work were presented in part at the International Conference on Health Policy Statistics, October 7–9, 2015, and the AcademyHealth Annual Research Meeting, June 26–28, 2016.*

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-020-05713-5) contains supplementary material, which is available to authorized users.*

Received July 1, 2018
Revised April 24, 2019
Accepted January 13, 2020
Published online March 19, 2020

KEY WORDS: observational studies; randomized controlled trials; propensity score; comparative effectiveness; benchmarking.

J Gen Intern Med 35(5):1396–1404
DOI: 10.1007/s11606-020-05713-5

INTRODUCTION

Randomized trials can provide valid evidence on the comparative effectiveness of interventions because randomization ensures the comparability of treatment groups in expectation and provides a "reasoned basis" for inference.[1] In addition to randomization, other aspects of randomized trials, such as blinding of investigators to treatment assignment, uniform follow-up protocols, and standardized procedures for outcome ascertainment, further enhance the ability of randomized trials to produce results that have a causal interpretation for the sample of randomized patients or for the (hypothetical) population from which the randomized patients can be viewed as a simple random sample (a property sometimes referred to as "internal validity"[2]). Unfortunately, because of high costs or ethical constraints, randomized trials are infeasible for many research questions and we have to rely on observational studies.[3] Furthermore, observational studies can help us better understand the effectiveness of interventions in routine clinical practice. Confounding poses a major threat to the validity of observational studies, but other problems, such as lack of blinding, variation in follow-up procedures, and non-standardized outcome ascertainment methods, can also invalidate observational study results.

Observational studies comparing interventions can be viewed as attempts to emulate target randomized trials[4]: other than the lack of randomization, observational studies can be designed to allow analyses that are similar to those conducted in pragmatic target trials. This view of observational studies motivates the use of concurrent control groups, new (incident)-user designs, and modern statistical methods for confounding control. For example, an increasing number of observational analyses use propensity score methods, that is, methods that rely on modeling the probability of treatment conditional on covariates. Because propensity score methods shift attention from modeling the outcome to modeling the treatment assignment (which is under investigator control in experimental studies), they are well suited to observational analyses emulating randomized trials.

When, for a given research question, evidence can be obtained from both observational analyses and large, well-conducted randomized trials, it may be possible to benchmark observational analysis methods by comparing their results against the randomized trial results.[5–8] If we find reasonable agreement in treatment effect estimates between sufficiently similar randomized trials and observational analyses across diverse clinical topics, we may be more willing to trust the findings of observational analyses on topics where randomized trials are unavailable. Benchmarking across multiple clinical topics can also reveal patterns that cannot be appreciated by examining each topic in isolation.[9–11]

In this paper, we synthesize the findings of systematic reviews comparing observational analyses using propensity score methods against randomized trials and explore whether intervention or outcome characteristics predict the degree of agreement between designs. On the basis of our findings and other relevant work, we propose steps towards more rigorous benchmarking of observational analyses in medicine.

METHODS

Search for Systematic Reviews

We searched PubMed from January 1, 2000, to April 30, 2017, using a combination of keywords and MESH terms to identify systematic reviews between observational analyses using propensity score methods (for short, "observational analyses") and randomized trials from high-impact general medical journals and journals known to publish methodological research; we provide our search strategy in the Appendix in the ESM. We also obtained a list of potentially relevant studies from the Scientific Resource Center Methods Library (AHRQ Effective Health Care Program; Portland VA Research Foundation; Portland, OR), a curated collection of citations related to methods of evidence synthesis and evidence-based medicine. The contents of the database are updated regularly using searches across multiple electronic databases and the gray literature (e.g., conference proceedings, technical reports, or dissertations). Finally, we identified potentially relevant studies by perusing the reference lists of eligible studies identified by our searches, Cochrane systematic reviews,[10,11] and our personal bibliographies.

Selection of Relevant Systematic Reviews

Both authors (SPF and IJD) independently screened titles and abstracts to identify relevant systematic reviews between observational analyses and randomized trials. SPF retrieved and examined the full texts of potentially eligible papers; IJD examined all excluded papers to verify that they did not meet the selection criteria. We selected English-language publications that reported "paired" comparisons of randomized trials versus observational studies using propensity score methods to evaluate the impact of the same interventions and comparators on similar binary or failure time outcomes. To be considered eligible, studies had to have focused on medical interventions and followed a systematic approach for identifying and selecting studies for their analyses. We required that studies had included at least three comparisons (i.e., at least three different intervention-comparator pairs examined by both observational studies and randomized trials) and focused on methodological issues related to study design and analysis. Finally, we excluded studies that did not report or allow the calculation of treatment effect estimates for each clinical topic.

Data Extraction

One reviewer extracted data from all eligible studies and a second reviewer verified them for accuracy. We collected the following information from each eligible study: clinical area; number of comparisons (i.e., combinations of interventions and outcomes) examined; number of studies of each design (randomized trials and observational studies); median number of participants by study design across comparisons; methods for identifying the studies contributing data; and methods for handling topics with multiple studies per design. We categorized different comparisons by whether the interventions examined were pharmacological or not; whether the outcomes examined were adverse or intended effects of treatment; and whether the outcome was death from any cause versus any other event. For each comparison, we extracted treatment effect estimates on the relative risk (RR) scale and their estimated sampling variance (we use "RR" as shorthand for odds, risk, or hazard ratios, as reported in the reviews). When data were only reported in graphs, we extracted numerical information with digitizing software (Engauge Digitizer; version 4.1[12]).

When multiple observational or randomized studies were available for the same topic, investigators used various approaches to obtain estimates for their comparisons. For example, when multiple randomized trials were paired with a single observational study, most studies performed a meta-analysis of the randomized trial results and compared the resulting summary estimate against the estimate from the observational study. To facilitate comparisons across reviews, whenever available, we used estimates from random effects meta-analyses.

Evidence Synthesis

We compared estimates from observational studies against estimates from randomized trials using similar interventions and outcomes in four ways: first, we examined whether the relative RR, that is, the ratio of the RR from the randomized trials over the RR from the observational studies *for each clinical topic*, was lower than 0.70 or greater than 1.43 (the reciprocal value), indicating "extreme" disagreement in the estimated magnitude of the effect;[13–15] we report analyses using different thresholds in the Appendix in the ESM. The coining of treatment effects varied *across clinical topics* and *across reviews,* and certain strategies for coining contrasts between designs can lead to bias when combining ratios of RRs across topics.[16] For this reason, and because we do not see a clear interpretation for the "pooled" relative RR, we did not combine estimates across topics. Second, for each clinical topic, we performed a test to compare estimates from both study designs, for the null hypothesis that the relative RR is significantly different from 1.[17] We defined two-sided *p* values < 0.05 as statistically significant. Third, we assessed whether estimates from different designs pointed to opposite directions of effect (i.e., one design showed benefit and the other harm). Fourth, we determined how often the randomized trial confidence interval included the point estimate from the observational study and vice versa. Last, we determined the observed and expected proportion of overlap of the 95% confidence intervals of randomized trial and observational analysis estimates (accounting for estimation uncertainty).[16]

We compared estimates from observational studies and randomized trials across all topics, separately for each review contributing data, and within specific subgroups defined by intervention and outcome characteristics. We conducted all analyses with Stata, version 14/IC (Stata Corp., College Station, TX).

RESULTS

Included Systematic Reviews

Figure 1 presents a summary of how we identified and selected relevant studies. Of the 2758 citations retrieved by our PubMed search, we deemed 190 potentially eligible. Searches in the Scientific Resource Center Methods Library, personal bibliographies, and perusal of reference lists yielded an additional 208 citations. In total, we examined 398 publications in full text. Of these, six studies, published between 2012 and 2016, met our inclusion criteria.[14,15,18–21]

We summarize the characteristics of included reviews in Table 1. In total, the reviews provided information on 127 comparisons in acute coronary syndrome care[14] (17 comparisons), diverse conditions requiring surgical intervention[18] (48 comparisons), critical care medicine and sepsis[15,19,20] (46 comparisons), and diverse clinical topics[21] (16 comparisons). Five out of six reviews exclusively considered death as the outcome of interest;[14,15,19–21] one review included outcomes other than death.[18] One review only used randomized trials published after the index observational studies had been conducted[21]; no other reviews considered the relative timing of publication of different designs as a selection criterion. The median sample size of randomized trials across reviews ranged from 118 to 985 participants; the median sample size of observational studies ranged from 433 to 5194 participants. All six reviews allowed for multiple studies per design and, when necessary, used meta-analysis methods to obtain design-specific pooled estimates, typically with random effects models. The reviews considered a median of 1 observational study (ranging from 1 to 14) and 2 randomized trials per topic (ranging from 1 to 21).

Comparing Effect Estimates Between Designs

Across all 127 comparisons, estimated RRs ranged from 0.11 to 4.01 (median 0.79) in observational studies and from 0.13 to 3.07 (median 0.86) in randomized trials. In general, randomized trial estimates were less precise than observational study estimates for the same topic, reflecting the generally larger sample sizes in observational studies. The treatment effect was statistically significantly different from the null value in 47 observational analyses (37%) and 14 randomized trials (11%). Figure 2 presents scatterplots of treatment effect estimates from the 127 topics, stratified by source review.

We summarize various measures of agreement across all 127 comparisons and stratified by source review in Table 2. The magnitude of the relative RR comparing randomized trials against observational studies was extreme (relative RR < 0.7 or > 1.43) in 68 of 127 (54%) topics; the percentage of extreme disagreements ranged from 35 to 69% across reviews. Disagreements were statistically significant in 12 of the 127 comparisons (9%) using a two-sided test; the percentage of statistically significant disagreements ranged from 4 to 25% across reviews. In 47 of the 127 comparisons (37%), estimates from observational studies and randomized trials pointed in the opposite directions of effect; the percentage of disagreements in direction from 18 to 50% across reviews. The randomized trial or observational study estimate was very close to 1 (between 0.95 and 1.05) in 13 of the 47 disagreements in the direction of effects; 1 of these disagreements was statistically significant. The randomized trial point estimate fell outside the confidence interval of the observational study in 43% of the comparisons and the frequency of non-coverage ranged from 31 to 59% across reviews. The observational study point estimate fell outside the confidence interval of the randomized trials in 22% of the comparisons and the frequency ranged from 13 to 50% across reviews. The 95% confidence intervals for each study design overlapped in 97% of the comparisons (123 of 127), which is close to the expected relative frequency of 99%.

Impact of Intervention and Outcome Characteristics on Agreement

Measures of agreement between designs stratified by intervention or outcome characteristics are summarized in Table 2. No clear pattern emerges: neither intervention type nor outcome characteristics were strongly associated with the degree of agreement between designs for any of the measures we considered (direction of effect, magnitude, or statistical significance). Inspection of Table 2 suggests that, if anything, the

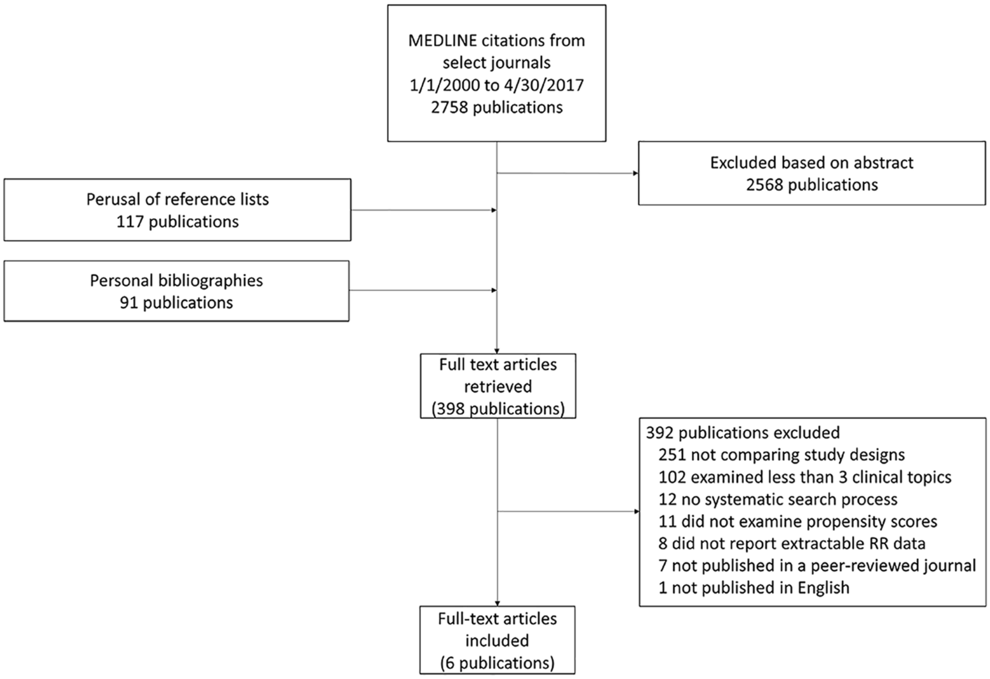
MEDLINE citations from select journals 1/1/2000 to 4/30/2017 2758 publications
Excluded based on abstract 2568 publications
Perusal of reference lists 117 publications
Personal bibliographies 91 publications
Full text articles retrieved (398 publications)
392 publications excluded
251 not comparing study designs
102 examined less than 3 clinical topics
12 no systematic search process
11 did not examine propensity scores
8 did not report extractable RR data
7 not published in a peer-reviewed journal
1 not published in English
Full-text articles included (6 publications)

Fig. 1 Diagram of the process for identifying relevant studies.

degree of agreement was more strongly associated with the source review than the intervention or outcome characteristics. The influence of the source review on the degree of agreement reflects the inherent differences among clinical areas, as well as the diversity of methodological approaches across reviews.

DISCUSSION

In our overview of 127 comparisons of observational analyses using propensity score methods against randomized trials, disagreements in effect estimates were not uncommon. Although rarely statistically significant, disagreements were sometimes large, did not appear predictable by treatment or outcome characteristics, and varied in frequency across reviews covering diverse clinical areas.

Benchmarking Observational Analysis Methods

Benchmarking observational analysis methods against completed randomized trials makes intuitive sense because observational studies comparing treatments can be viewed as attempts to emulate pragmatic target trials.[4,22–27] This kind of benchmarking has a long history in medicine[5,6,11,28,29] and other fields that rely on observational data to draw causal inferences.[30–37] However, previous reviews of benchmarking attempts that used published data have not focused on modern methods for confounding control (e.g., only 2 of 15 studies in a recent Cochrane review used propensity score methods[11]). We focused our analysis on studies using propensity score methods because they are a natural choice when viewing observational analyses as attempts to emulate target trials: the propensity score is known in randomized trials but typically has to be modeled and estimated in observational studies. Our approach to benchmarking, however, applies to other causal inference methods provided they can be conceptualized as components of a target trial emulation (e.g., some difference-in-difference or instrumental variable analyses).

The seminal comparisons of observational econometric analyses against large social experiments, first published in the 1980s, are of particular relevance to our work.[30,31] These investigations showed that different observational analyses of the same data produced different results between them and compared with the social experiments. Recent analyses of the same data with propensity score methods suggested that agreement between designs can be improved by careful participant selection and choice of covariates for confounding control,[38] but even with these improvements, agreement remained sensitive to modeling choices.[39–41]

Interpretation of Benchmarking Results

The interpretation of comparisons between observational studies and randomized trials is challenging because disagreements between designs can occur for many reasons, even when studies are conducted in accordance with rigorous research standards.[4,42,43] First, disagreements are expected when different designs examine different interventions or outcomes, as when observational studies use broader intervention definitions and less systematic outcome ascertainment methods compared with randomized trials, or when studies ascertain outcomes at different time points. Second, estimates of population-averaged

Table 1 Characteristics of Included Reviews Comparing Observational Studies Using Propensity Score Methods with Randomized Trials

Systematic comparison	Topic area	Number of comparisons (number of observational studies/randomized trials)	Median number of patients in observational studies/ randomized trials	Sources of observational studies	Sources of randomized trials	Pairing of observational studies and randomized trials	Methods for handling topics with multiple studies per design
Dahabreh (2012)[14]	Acute coronary syndromes	17 (21/63)	5194/177	MEDLINE (top 8 journals in “Cardiac and Cardiovascular systems” and top 4 journals in “Medicine, general and internal” by impact factor)	CDSR, MEDLINE, evidence-based guidelines of the American Heart Association/ American College of Cardiologists, a compendium of medical therapeutics (Washington Manual of Medical Therapeutics), and reference lists	Two reviewers matched each observational study to at least one randomized trial, using a PICO scheme and a structured search protocol	Meta-analysis using random effects models (DerSimonian-Laird); common effect models (sensitivity analysis)
Lonjon (2014)[18]	Surgery	48 (70/94)	2049/179	MEDLINE via PubMed	MEDLINE via PubMed	For each eligible observational study, a single reviewer searched for matching randomized trials using a PICO scheme	Meta-analysis using random effects models (DerSimonian-Laird); common effect models
Zhang (2014a)[19]	Sepsis	8 (14/40)	4641/359*	PubMed, Scopus, EBSCO	PubMed	Two investigators independently matched observational studies to randomized trials or most updated systematic review of randomized trials using a PICO scheme	Meta-analysis using random effects models (DerSimonian-Laird)
Zhang (2014b)[20]	Critical care medicine	20 (20/130†)	433/150	PubMed	PubMed	Each observational study was matched to one or more randomized trials using PICO	Meta-analysis using random effects models (DerSimonian-Laird)
Kitsios (2015)[15]	Critical care medicine	18 (21/58)	1327/118	MEDLINE (top 5 journals that publish primary clinical research studies and top 4 journals in “Medicine, general and internal” by total citations), reference lists	CDSR, PubMed, reference lists, and reference files of intensivists on review team	Two reviewers independently matched each observational study to at least one randomized trial, using a PICO scheme and a structured search protocol	Meta-analysis using random effects models (REML)
Hemkens (2016)[21]	Mixed clinical areas	16 (16/36)	2086/985	PubMed	PubMed, Cochrane Library	Searched for observational studies published before an randomized trial was conducted, up to 2010; then, searched for randomized trials, systematic reviews, or meta-analyses of trials conducted after 2010 on the same clinical topics	Meta-analysis using random effects models; common effect models (sensitivity analysis); Peto’s method for event rates $<1\%$

CDSR, Cochrane Database of Systematic Reviews; EBSCO, Elton B. Stephens Co. database; NRS, nonrandomized studies; PICO, population, interventions/comparators, outcomes

**Average study size, by design*

†Exact number of randomized trials not specified so estimate is used

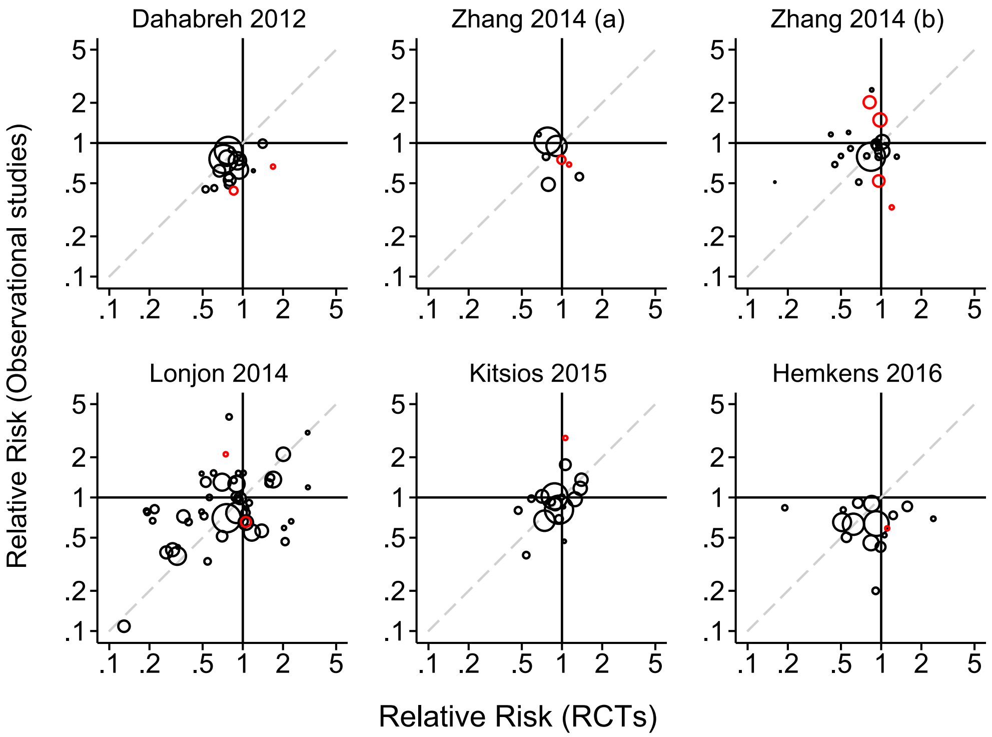

Fig. 2 Scatterplot of treatment effect estimates from randomized trials and comparable observational studies. The scatterplots titled Zhang 2014 (a) and Zhang 2014 (b) correspond to references [19] and [20], respectively. Red (black) circles represent statistically significant (non-significant) differences.

(marginal) effects will vary across designs when effect modifiers are differentially distributed in the underlying populations. This is a concern for the comparisons we examined, because commonly used propensity score–based methods and unadjusted between-group comparisons in randomized trials estimate marginal effects. Third, different observational analyses estimate different causal parameters, which can lead to disagreements with randomized trials even in the absence of bias or differences in underlying populations. For example, outcome regression with the propensity score as a covariate does not in general estimate the same causal parameter as unadjusted treatment-group comparisons in randomized trials. Fourth, biases can affect both observational studies and randomized trials and can produce large disagreements. Baseline confounding by unmeasured variables in observational studies is the leading explanation when disagreements occur, because such confounding can never be ruled out in observational studies but is negligible in large well-conducted randomized trials.[44] But other biases, such as selection bias (e.g., differential loss-to-follow-up) and measurement error bias, affect both designs and can induce large disagreements.[45] Finally, disagreements can arise by chance, particularly when studies are small and outcomes are rare.

To some extent, all these mechanisms inducing disagreements between observational analyses and randomized trials affected the systematic reviews we examined. The observational analyses used data from diverse sources and were compared against published randomized trial results; studies were conducted independently in different study populations; and many studies included in the reviews had small sample sizes. Studies of different designs were matched on the basis of limited information available in published reports and no attempt was made to standardize causal contrasts or approaches for bias control. Under these conditions, close matching of populations, interventions, and outcomes is near impossible and randomized trials provide an imperfect reference standard.[46] We conjecture that the results we reviewed *place a lower bound on the degree of agreement* possible between observational analyses and randomized trials: closer matching of populations, richer data, and use of modern methods for addressing biases in observational studies[43] and randomized trials should improve agreement. Of note, our conjecture rests on the assumption that selective reporting of observational analyses and randomized trials has been limited or independent of the direction and magnitude of disagreements. If, for example, publications of observational studies with results in strong disagreement with randomized trials have been suppressed, then the studies included in our review do not fully reflect the true magnitude of disagreements between designs.

Towards Better Benchmarking

We found that disagreements between observational studies using propensity score methods and randomized trials were not infrequent and the available data could not be used to

Table 2 Comparisons of Results from Randomized Controlled Trials and Observational Studies in the Included Systematic Reviews

Subgroups	Number of comparisons	Point estimates in opposite directions (%)	Relative RR < 0.70 or > 1.43 (%)	Statistically significant discrepancy (%)	Randomized trial point estimate outside observational CI (%)	Observational point estimate outside randomized trial CI (%)	Observed randomized trial and observational CI overlap (%)	Expected randomized trial and observational CI overlap (%)
All comparisons	127	47 (37)	68 (54)	12 (9)	55 (43)	28 (22)	123 (97)	125.43 (99)
Source of data								
Dahabreh (2012)	17	3 (18)	6 (35)	2 (12)	10 (59)	3 (18)	15 (88)	16.75 (99)
Lonjon (2014)	48	22 (46)	28 (58)	2 (4)	15 (31)	8 (17)	48 (100)	47.44 (99)
Zhang (2014a)	8	4 (50)	4 (50)	2 (25)	3 (38)	4 (50)	8 (100)	7.93 (99)
Zhang (2014b)	20	8 (40)	12 (60)	4 (20)	10 (50)	8 (40)	18 (90)	19.75 (99)
Kitsios (2015)	18	5 (28)	7 (39)	1 (6)	8 (44)	3 (17)	18 (100)	17.81 (99)
Hemkens (2016)	16	5 (31)	11 (69)	1 (6)	9 (56)	2 (13)	16 (100)	15.75 (98)
Intervention type								
Both drugs	38	13 (34)	17 (45)	6 (16)	16 (42)	14 (37)	37 (97)	37.57 (99)
At least one not a drug	89	34 (38)	51 (57)	6 (7)	39 (44)	14 (16)	86 (97)	87.86 (99)
Effect of treatment								
Adverse effect	12	6 (50)	6 (50)	1 (8)	5 (42)	2 (17)	12 (100)	11.85 (99)
Intended effect	115	41 (36)	62 (54)	11 (10)	50 (43)	26 (23)	111 (97)	113.57 (99)
Outcome type								
Death from any cause	102	37 (36)	58 (57)	11 (11)	45 (44)	23 (23)	98 (96)	100.70 (99)
Other outcomes	25	10 (40)	10 (40)	1 (4)	10 (40)	5 (20)	25 (100)	24.72 (99)

Percentages are calculated for each comparison source or type. CI confidence interval; RR relative risk discern the reasons behind the disagreements. We propose that better benchmarking can improve observational analyses, including those that use propensity score methods, and help identify under what conditions observational analyses can produce valid results that can inform clinical decisions.

Recent work in education shows how benchmarking can work under ideal conditions: in a doubly randomized preference design, undergraduate students were randomly assigned to participate in a randomized trial or an observational study.[47] Students allocated to the randomized trial were further randomized to training in mathematics versus vocabulary; those allocated to the observational study were allowed to self-select into the same programs. Analyses in the observational study, including analyses using propensity score methods, produced effect estimates similar to those in the randomized trials, when adjusting for a rich set of covariates, but not when adjusting only for "predictors of convenience."[47,48] An independent replication has largely confirmed these findings.[49] This evidence from doubly randomized preference designs illustrates that valid causal inference *is possible* in observational studies, provided all important confounding variables are measured. These designs, however, pose substantial logistical and ethical challenges in medicine and cannot be conducted at scale. At the other extreme, the wide availability of routinely collected data has made it possible to carry out observational analyses with little input from human experts.[29] The usefulness of automated large-scale evaluations of observational analysis methods, however, is limited because the complex methodological decisions needed to conduct clinically relevant observational analyses cannot be reduced to a small set of predefined options.[50]

Better benchmarking is possible using observational analyses that explicitly attempt to emulate target trials in diverse areas where strong background knowledge on the direction and magnitude of effects is available (e.g., from large multicenter pragmatic trials).[51] Such studies should be conducted across diverse topics, using multiple data sources, by teams combining clinical, epidemiological, and statistical expertise. To eliminate predictable causes of disagreements, comparisons of observational studies using against randomized trials should take advantage of data from cohort studies where a subset of participants are randomized and others self-select into treatment[52,53] or pragmatic trials embedded in healthcare systems and registries;[54,55] focus on well-defined causal contrasts; and use methods to control biases that affect both designs. Complete reporting of emulation attempts can be encouraged by prospective registration.

Acknowledgments: *The authors thank Ms. Robin Paynter, MLIS (Scientific Resource Center; AHRQ Effective Health Care Program; Portland VA Research Foundation; Portland, OR), for the help with retrieving citations from the Methods Library of the Scientific Resource Center.*

Corresponding Author: *Issa J. Dahabreh, MD ScD; Center for Evidence Synthesis in Health, Brown University School of Public Health, Providence, USA (e-mail: issa_dahabreh@brown.edu).*

Funding Information *This work was supported by two Patient-Centered Outcomes Research Institute (PCORI) Methods Research Awards (ME-130603758 and ME150227794 to IJD).*

Data Availability *The data for this study are provided as a supplementary material to this paper.*

Compliance with Ethical Standards:

Ethics Approval: *Not required for this study.*

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Disclaimer: *All statements in this paper, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of Governors, or the Methodology Committee. Additionally, the funder was not involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.*

REFERENCES

1. **Fisher RA**. The design of experiments: Oliver And Boyd; Edinburgh; London, 1937.
2. **Shadish WR**, **Cook TD**, **Campbell DT**. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin, 2001.
3. **Dahabreh IJ**. Randomization, randomized trials, and analyses using observational data: A commentary on Deaton and Cartwright. Soc Sci Med 2018;210:41–44.
4. **Hernan MA**, **Robins JM**. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016;183(8):758–64.
5. **Concato J**, **Shah N**, **Horwitz RI**. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342(25):1887–92.
6. **Benson K**, **Hartz AJ**. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342(25):1878–86.
7. **MacLehose RR**, **Reeves BC**, **Harvey IM**, **et al**. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess 2000;4(34):1–154.
8. **Ioannidis JP**, **Haidich AB**, **Pappa M**, **et al**. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286(7):821–30.
9. **Deeks JJ**, **Dinnes J**, **D'Amico R**, **et al**. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii-x, 1–173.
10. **Kunz R**, **Vist G**, **Oxman AD**. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev 2007(2):MR000012.
11. **Anglemyer A**, **Horvath HT**, **Bero L**. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014(4):MR000034.
12. **Mitchell M.** Engauge Digitizer. A free open-source software to extract data points from a graph image. Hosted on SourceForge at: http://digitizer.sourceforge.net; 2002.
13. **Sterne JA**, **Jüni P**, **Schulz KF**, **et al**. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21(11):1513–24.
14. **Dahabreh IJ**, **Sheldrick RC**, **Paulus JK**, **et al**. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J 2012;33(15):1893–901.
15. **Kitsios GD**, **Dahabreh IJ**, **Callahan S**, **et al**. Can We Trust Observational Studies Using Propensity Scores in the Critical Care Literature? A Systematic Comparison With Randomized Clinical Trials. Crit Care Med 2015;43(9):1870–9.
16. **Franklin JM**, **Dejene S**, **Huybrechts KF**, **et al**. A Bias in the Evaluation of Bias Comparing Randomized Trials with Nonexperimental Studies. Epidemiol Methods 2017;**6**(1).

17. **Altman DG**, **Bland JM**. Interaction revisited: the difference between two estimates. BMJ 2003;326(7382):219.
18. **Lonjon G**, **Boutron I**, **Trinquart L**, **et al**. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg 2014;259(1):18–25.
19. **Zhang Z**, **Ni H**, **Xu X**. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis? J Crit Care 2014;29(5):886 e9–15.
20. **Zhang Z**, **Ni H**, **Xu X**. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. J Clin Epidemiol 2014;67(8):932–9.
21. **Hemkens LG**, **Contopoulos-Ioannidis DG**, **Ioannidis JP**. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ 2016;352:i493.
22. **Cochran WG**. Planning and analysis of observational studies: Wiley, Hoboken 2009.
23. **Robins J.** A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 1986;7(9–12):1393–512.
24. **Miettinen OS**. The clinical trial as a paradigm for epidemiologic research. J Clin Epidemiol 1989;42(6):491–6; discussion 97-8.
25. **Rubin DB**. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26(1):20–36.
26. **Hernan MA**, **Alonso A**, **Logan R**, **et al**. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008;19(6):766–79.
27. **Rosenbaum PR**. Design of observational studies: Springer, Berlin 2010.
28. **Kunz R**, **Oxman AD**. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317(7167):1185–90.
29. **Ryan PB**, **Madigan D**, **Stang PE**, **et al**. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med 2012;31(30):4401–15.
30. **LaLonde RJ**. Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 1986:604–20.
31. **Fraker T**, **Maynard R**. The adequacy of comparison group designs for evaluations of employment-related programs. J Hum Resour 1987:194–227.
32. **Lipsey MW**, **Wilson DB**. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. Am Psychol 1993;48(12):1181–209.
33. **Glazerman S**, **Levy DM**, **Myers D**. Nonexperimental versus experimental estimates of earnings impacts. Ann Am Acad Pol Soc Sci 2003;589(1):63–93.
34. **Agodini R**, **Dynarski M**. Are experiments the only option? A look at dropout prevention programs. Rev Econ Stat 2004;86(1):180–94.
35. **Michalopoulos C**, **Bloom HS**, **Hill CJ**. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? Rev Econ Stat 2004;86(1):156–79.
36. **Hill JL**, **Reiter JP**, **Zanutto EL**. A comparison of experimental and observational data analyses. Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family 2004:49–60.
37. **Cook TD**, **Shadish WR**, **Wong VC**. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. J Policy Anal Manag 2008;27(4):724–50.
38. **Dehejia RH**, **Wahba S**. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. J Am Stat Assoc 1999;94(448):1053–62.
39. **Smith JA**, **Todd PE**. Reconciling conflicting evidence on the performance of propensity-score matching methods. Am Econ Rev 2001;91(2):112–18.
40. **Smith JA**, **Todd PE**. Does matching overcome LaLonde's critique of nonexperimental estimators? J Econ 2005;125(1):305–53.
41. **Dehejia R.** Practical propensity score matching: a reply to Smith and Todd. J Econ 2005;125(1):355–64.
42. **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(2):217–30.
43. **Franklin JM**, **Schneeweiss S**. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;102(6):924–33.
44. **Greenland S**, **Robins JM**. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986;15(3):413–9.
45. **Hernan MA**, **Sauer BC**, **Hernandez-Diaz S**, **et al**. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;79:70–75.
46. **Berk RA**. Randomized experiments as the bronze standard. J Exp Criminol 2005;1(4):417–33.
47. **Shadish WR**, **Clark MH**, **Steiner PM**. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J Am Stat Assoc 2008;103(484):1334–44.
48. **Steiner PM**, **Cook TD**, **Shadish WR**, **et al**. The importance of covariate selection in controlling for selection bias in observational studies. Psychol Methods 2010;**15**(3):250–67.
49. **Pohl S**, **Steiner PM**, **Eisermann J**, **et al**. Unbiased causal inference from an observational study: Results of a within-study comparison. Educ Eval Policy Anal 2009;**31**(4):463–79.
50. **Gruber S**, **Chakravarty A**, **Heckbert SR**, **et al**. Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug-outcome association. Pharmacoepidemiol Drug Saf 2016;25(9):973–81.
51. **Dahabreh IJ**, **Kent DM**. Can the learning health care system be educated with observational data? JAMA 2014;312(2):129–30.
52. **Olschewski M**, **Scheurlen H**. Comprehensive Cohort Study: an alternative to randomized consent design in a breast preservation trial. Methods Inf Med 1985;24(3):131–4.
53. **Schmoor C**, **Olschewski M**, **Schumacher M**. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. Stat Med 1996;15(3):263–71.
54. **Califf RM**, **Robb MA**, **Bindman AB**, **et al**. Transforming Evidence Generation to Support Health and Health Care Decisions. N Engl J Med 2016;375(24):2395–400.
55. **Li G**, **Sajobi TT**, **Menon BK**, **et al**. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol 2016;80:16–24.