REVIEW

For reprint orders, please contact: reprints@futuremedicine.com

Journal of Comparative Effectiveness Research

Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go?

Bijan J Borah*[1,2], James P Moriarty[2], William H Crown[3] & Jalpa A Doshi[4]

[1]Mayo Clinic Medical School & the Division of Health Care Policy & Research, Mayo Clinic, Rochester, MN, USA
[2]Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA
[3]Optum Labs, Cambridge, MA, USA
[4]Division of General Internal Medicine, University of Pennsylvania, Philadelphia, PA, USA
*Author for correspondence:
Tel.: +1 507 284 2873
Fax: +1 507 284 1731
borah.bijan@mayo.edu

Propensity score (PS) methods have proliferated in recent years in observational studies in general and in observational comparative effectiveness research (CER) in particular. PS methods are an important set of tools for estimating treatment effects in observational studies, enabling adjustment for measured confounders in an easy-to-understand and transparent way. This article demonstrates how PS methods have been used to address specific CER questions from 2001 through to 2012 by identifying six impactful studies from this period. This article also discusses areas for improvement, including data infrastructure, and a unified set of guidelines in terms of PS implementation and reporting, which will boost confidence in evidence generated through observational CER using PS methods.

KEYWORDS: comparative effectiveness research ▪ general and internal medicine ▪ high impact ▪ impact factors ▪ observational studies ▪ policy implications ▪ propensity score methods

Defined as the "…research evaluating and comparing health outcomes and clinical effectiveness, risks and benefits of two or more medical treatments, services and items," the Patient Protection and Affordable Care Act (ACA) of 2010 formalized comparative effectiveness research (CER) to be an important step in the US federal government's efforts to improve the quality of healthcare [101]. The term CER encompasses both comparative effectiveness and safety research, and although CER was being conducted in different forms in the fields of epidemiology, health economics and outcomes research and health services research even before the ACA, what the law did was to explicitly recognize that CER can be used to determine the most effective and/or safe intervention from among the existing alternatives in the real-world clinical practice setting [1].

Randomized controlled trials (RCTs) have the strongest research design for assessing treatment effects because the treatment groups are balanced on both observed and unobserved covariates. This addresses the bias problem (e.g., confounding by indication or channeling bias or selection bias, among others) that is common in real-world observational studies and is introduced by unobserved covariates that are correlated with both the treatment variable and outcomes. While RCT designs are the gold standard for evaluating comparative 'efficacy' of two or more medical interventions, there are several reasons why RCTs may not be suitable for generating robust evidence of comparative 'effectiveness'. Deriving effectiveness from RCTs may not be straightforward due to a potential lack of external validity of the RCTs – the controlled setting of an RCT often may not occur in real-life [2]. Furthermore, due to their often restrictive enrollment criteria, patients included in RCTs are not generally representative of real-world clinical practice patients [2].

Future Medicine part of fsg

doi:10.2217/CER.13.89 © BJ Borah, □
JP Moriarty, WH Crown & JA Doshi

ISSN 2042-6305

As a result, 'pragmatic' randomized trials for conducting CER have been proposed in order to address the issues of generalizability of traditional RCTs [3,4]. However, it may not be feasible even to conduct pragmatic RCTs for all of the CER questions. RCTs are time consuming and expensive to design and conduct, particularly for comparing the long-term effectiveness or safety of interventions wherein patients have to be followed for an extended period of time or for RCTs involving rare outcomes that require sample sizes of hundreds and thousands of patients [2]. On the other hand, observational study designs often offer the advantages of relatively quick and inexpensive analyses of real-world data on large number of patients that are frequently available from existing databases with longitudinal follow-up over several years. As a result of the pressing need for CER evidence, in combination with improving data availability and richness, it is very likely that the demand for evidence from observational studies will continue to grow, despite their inherent limitation of different biases that these studies might induce.

The contribution of observational studies to the field of medicine is substantial [5,6]. Well-designed observational studies controlling for selection or other biases have been found to yield similar results as RCTs [7–11]. At the same time, many of the previous findings based on observational studies have been contradicted by RCTs, thus raising questions about the credibility of observational studies (e.g., the protective effects of hormone therapy on myocardial infarction [MI] and β-carotene on lung cancer, among others) [12,13]. Many of the divergent findings between RCTs and observational studies were attributed to the differences in enrolled patient populations and the specific methods used to account for inherent biases in observational studies [5,12,13]. Thus, it is crucial that appropriate statistical methods be used to analyze such data, without which drawing reliable statistical conclusions from observational studies can be hindered.

Propensity score (PS) methods have emerged as an important set of techniques in the statistical and analytic toolkit for studies involving observational data. The use of PS methods in observational studies (regardless of their focus on CER) published in the medical literature has proliferated in recent years. A total of 4835 citations were identified for the search term "propensity scor*" between the calendar years 1987 and 2012 in PubMed. The number of citations increased exponentially over the last decade, from 42 in 2002 to 1118 in 2013 [102]. Furthermore, numerous systematic reviews of PS-based studies have been published; however, the majority of these studies have focused on highlighting methodological issues in the applications of PS methods in published studies [7,10,14–18].

This review has three goals. First, to examine the trends in PS methods-based observational CER studies published in high-impact-factor journals (used as a proxy for identifying studies of high quality and impact). Second, to provide an overview of selected highly cited applications of PS methods with a focus on the key policy, clinical or research implications emerging from the results of these observational CER studies. The aim is to present a few examples in order to highlight the role that PS methods has played in evaluating and informing several important CER questions. Third, we conclude with suggestions for improvements to boost confidence in evidence generated through observational CER using PS methods.

Brief overview of PS methods & their advantages over traditional regression models

A PS is the conditional probability of receiving the treatment/intervention under study given the observed characteristics. PS methods facilitate balance in the distribution of observed covariates between the treatment and comparison groups. Under the assumption that treatment assignment depends only on observed characteristics ('selection on observables' assumption), Rosenbaum and Rubin showed that balancing on the PS is equivalent to balancing on the background covariates [19]. Estimated PSs can be used in many ways to achieve balance between the treatment and control cohorts, including matching [19,20], stratification [19,21,22], covariate adjustment [19,21,23] and inverse PS adjustment [24,25]. Additionally, extensions of the above methods are also available, including doubly robust estimation techniques [26] and multidimensional PS techniques [27], which, among other things, enable the integration and automation of PS-based confounder adjustment modeling by incorporating the effect of time-varying covariates. There are many aspects to implementing PS methods, including assessment of the goodness-of-fit of the model used to estimate PSs, as well as assessment of the balance between covariates and multivariate adjustment following the application of PS methods in order to account for residual bias [21,23,28–32]. Note that the goodness-of-fit of the PS model needs to be discussed in the context of how well it balances the observed covariates because it is the latter that reflects the success of the PS method [2,29,30]. As with conventional multivariate methods that adjust for only measured confounders, PS methods also account for only measured confounders. Several studies have found similar estimates of treatment effects between PS-based and conventional regression-based studies [17,18]. However, there are certain advantages of using PS methods over conventional multivariate regression methods:

- While assessing the outcomes between two matched cohorts, it is rather straightforward to draw inferences from the matched analyses [19];
- Conventional regression methods do not ensure that the groups being compared have a common support (overlapping covariate distributions). Treatment effect estimates obtained from conventional regression methods outside of the range of common support are not generally valid, except under strong assumptions [16]. Matching is particularly good for addressing the common support issue. Stratification also helps in this regard, although in the extremes of the strata, there may still be observations that lie outside the bounds of common support;
- In situations with rare outcomes with multiple confounders [16,33,34], PS methods offer a clear advantage over traditional regression methods;
- When key covariates are missing or omitted, which is a common problem in observational database analyses, both PS and common prognostic models (e.g., regression methods) yield similar results, but the potential bias due to a misspecified response model can be much larger than the bias that can be introduced by similar misspecification in PS model [35];
- In CER, while comparing the effectiveness of two drugs, for example, confounding by indication is a significant threat to validity; however, unlike conventional regression models, PS methods circumvent this limitation either by explicitly modeling the indications for use or nonuse of the drugs under study or by ensuring that the postmatch distribution of the PS between treated and control subjects has common overlap, so that patients in the nonoverlapping areas are those with a very high or very low indication for the drug usage [16]. In addition, as Glynn *et al.* showed, the heterogeneity of treatment effects by the strength of indication can also be discovered using a PS stratification method, which is not feasible when using conventional methods [16];
- An important prerequisite for the widespread translation of observational CER results is the robustness of the findings to specific study designs (i.e., the outcomes of the study must not vary by the specific study design that is adopted). Such objectivity in observational CER can be achieved by PS methods, which can help construct the intervention and control cohorts with a similar distribution of baseline characteristics without having access to data on study outcomes [36,37];
- PS methods afford the application of regression calibration approaches in order to correct for measurement errors in certain observed covariates [16,38]. For example, it may be possible to collect more detailed/complete information on a subpopulation of the main study population, which enables the estimation of the 'gold standard' PS, which can then be used to correct for the measurement error in the main study [38,39].

Literature search & study selection strategy

Criteria for selection of studies to examine trends in PS methods use in observational CER studies

Potential studies were identified in Scopus, a multidisciplinary citation index covering 20,000 peer-reviewed journals in biomedicine and health, science, technology and the arts and humanities from 1995 to the present [103]. All available fields were searched for the phrase "propensity scor*". The * truncation symbol indicates any string of text is acceptable such as 'score', 'scores' and 'scoring', among others. This searching approach within Scopus was sensitive in finding all the studies that had the phrase "propensity scor*" within their texts; however, it lost specificity by including articles that might have the phrase only in the reference section, but did not use this methodology itself. To account for this limitation, specific full-text exclusionary criteria were designed in order to appropriately exclude articles that only had "propensity scor*" appear in the reference section. These criteria were also designed to exclude studies using PS methods, but not in the context of CER.

The search was limited to the years 2001 through to 2012. Within each year, the journals searched were restricted to the top five general and internal medicine journals publishing original research. Ranking of journals was based on impact factors for the general and internal medicine category in each year's journal citation reports [104].

■ Criteria for selection of studies for review

All articles meeting the Scopus search criteria (described above) were categorized into the following 2-year groupings: 2001–2002, 2003–2004, 2005–2006, 2007–2008, 2009–2010 and 2011–2012. Since impact factors for journal articles may be higher the earlier the article was published, we wanted to ensure that we picked impactful studies from each of the above 2-year periods. Full texts were obtained for the five articles within each 2-year grouping having the highest number of citations in Scopus. These 30 articles then underwent a full-text review. Articles needed to meet the following criteria from the full-text review to be considered in the analysis:

- Propensity analysis involves comparison of at least two distinct nonplacebo pharmaceutical drugs, medical/surgical procedures or medical devices. If 'standard of care' as a nonplacebo comparator is used in the study, it must be explicitly described. Placebo as a comparator is ruled out since CER is expected to provide comparative evidence between alternative treatment options that will be most relevant to the decision-maker in the real-world setting [1];
- Treatments or interventions must be distinct, but not simply being differing intensities of the same intervention;
- Must be original work. Review articles, theoretical articles or papers on PS methods other than their applications and editorials were excluded.

The review of the full-text articles was conducted by two of the authors. Complete agreement between the two reviewers was required for final inclusion of an article. Discordance on study inclusion was discussed by the reviewers until arriving at a consensus. Initial review resulted in 80% agreement on the articles to be included for analysis. For the discordant articles, consensus was reached following minimal discussion between the reviewers.

Trends in the applications of PS methods in observational CER studies

Our initial search using the Scopus criteria "propensity scor*" returned 273 CER articles in the top five general and internal medicine journals between 2001 and 2012. Figure 1 shows the number of CER articles that used PS methods between the time period from 2001 to 2012 for the five journals in question. The graph indicates an upward trend in the publications of PS-based studies in the top five general and internal medicine journals in the last 3 years.

Overview of selected studies

Details regarding the 30 articles selected for a full-text review, which included the top five articles in each 2-year period in terms of number of citations, are provided in Supplementary Table 1 (see online at www.futuremedicine.com/doi/suppl/10.2217/cer.13.89). These articles came from the following journals: *Annals of Internal Medicine* (n = 3), *The Journal of the American Medical Association* (n = 15), *The Lancet* (n = 2) and *The New England Journal of Medicine* (n = 10). Supplementary Table 2 documents which of the three inclusion criteria were satisfied by the 30 articles that were eligible for full-text review. Upon applying the full-text review criteria, six out of the 30 articles were considered eligible for inclusion in our study, which came from the following journals: *The Lancet* (n = 1), *The Journal of the American Medical Association* (n = 2) and *The New England Journal of Medicine* (n = 3) [40–45]. Three of the six studies focused on comparative safety outcomes, whereas the other three focused on comparative effectiveness outcomes. The following sections provide a brief overview of each of these six studies.

■ Study 1: Wang *et al.* (2005)

In 2008, the US FDA issued a public advisory to patients and providers on the potential safety concerns associated with conventional antipsychotics prescribed to the elderly for dementia-related psychosis, and also required manufacturers to add a boxed warning regarding the mortality risk for this class of drugs [105]. Although the FDA cited two different observational studies in this advisory [46,47], the study by Wang *et al.* arrived at similar conclusions 3 years earlier regarding the comparable risk profile of both conventional and atypical antipsychotic medications [40]. Their study used a propensity-matched sample of elderly patients with a drug insurance benefit from the state of Pennsylvania between 1994 and 2003, and found that conventional antipsychotic medications have similar risks as those of atypical antipsychotic medications [40]. This study used PS-adjusted Cox proportional hazard models to confirm the findings from traditional multivariate regression models [40].

The results of this study highlight how observational CER using PS methods was able to provide very early evidence of the comparative safety of medications in elderly patients with dementia-related psychosis – a vulnerable patient group that is most often excluded from traditional RCTs. Upon further confirmatory evidence, the findings of this study had direct implications for the generation of an FDA advisory.

■ Study 2: Lagerqvist *et al.* (2007)

The study by Lagerqvist *et al.* sets out to evaluate the hypothesis that drug-eluting stents (DESs) were associated with a higher rate of long-term adverse outcomes (deaths and MI) compared with bare-metal stents (BMSs) [41]. Using national registry data from Sweden, their study found that DESs were associated with a higher risk of death and a higher risk of the composite outcome of death or MI (Table 1) [41]. They used the estimated PS as a covariate in the multivariate adjustment through Cox modeling [19,41].

The results of this study provided supporting evidence for the need for a large RCT in order to further validate these findings. RCT evidence that has accumulated since the publication of this study has suggested the existence of a comparable safety and effectiveness profile between DESs and BMSs [48].

■ Study 3: Eisenstein *et al.* (2007)

At approximately the time that Eisenstein *et al.* undertook their study [42], some new evidence suggested that a shorter regimen of clopidogrel therapy was associated with a higher incidence of deaths and/or MI in patients undergoing percutaneous coronary intervention with DESs and BMSs [49,50]. This observational study of patients receiving DESs or BMSs from a single institution, which used inverse PS-weighted methods [19,24,51,52], along with multivariate Cox regression, showed that extended clopidogrel use in DES patients was associated with a reduced risk of death and the composite of death or MI.

The results of this observational CER study highlighted to the clinical community that clopidogrel use appeared to have benefited patients undergoing DES treatment; however, the duration of its use needs to be confirmed through a large RCT.

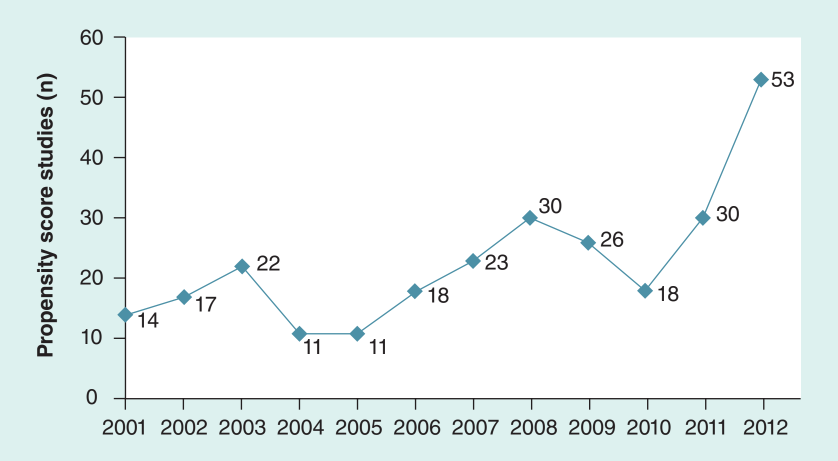

Figure 1. Number of comparative effectiveness research studies using propensity scores in the top five general and internal medicine journals between 2001 and 2012.

■ Study 4: O'Donoghue *et al.* (2009)

Using PS methods, O'Donoghue *et al.* found that concomitant use of proton pump inhibitors (PPIs) was not suggestive of attenuating the effects of clopidogrel versus prasugrel [43]. Although the data for their study came from two separate RCTs – PRINCIPLE-TIMI 44 [53] and TRITON-TIMI 38 [54] – PPI use was at the physician's discretion, which necessitated the adjustment of patient baseline characteristics through PS methods. In the PRINCIPLE-TIMI 44 trial, the primary outcome of the inhibition of platelet aggregation was assessed between percutaneous coronary intervention patients receiving clopidogrel or prasugrel with and without PPIs [53]. In the TRITON-TIMI 38 trial, acute coronary syndrome patients were randomized to receive clopidogrel and prasugrel, of which 33% of the patients were on PPIs at the time of randomization, where the primary outcome of interest was the composite of cardiovascular death, MI or stroke (Table 1) [54]. O'Donoghue *et al.* used PS with stratification, in which multivariate adjustment through Cox regression was conducted within each of the 11 strata, and the final treatment effect of PPIs on the outcome was derived by averaging across the strata treatment effects [43].

This study was a novel application of data collected in clinical trials in order to answer a CER

Table 1. Summary of the selected studies reviewed in this article.

Study (year)	Comparative effectiveness or safety study	Medical specialty	Data source	Treatment	Comparator	Outcome
Wang *et al.* (2005)	Comparative safety	Psychiatry	Pennsylvania Pharmaceutical Assistance Contract for the Elderly, a large state prescription-benefits program for the elderly in the USA (1994–2003)	Conventional antipsychotic medication	Atypical antipsychotic medication	Death in the following time intervals: ≤180 days, <40 days, 40–79 days, 80–180 days
Lagerqvist *et al.* (2007)	Comparative safety	Cardiology	Swedish Coronary Angiography and Angioplasty Registry (2003–2004)	DESs	BMSs	Primary: composite of death or MI Secondary: death, MI, revascularization and restenosis
Eisenstein *et al.* (2007)	Comparative effectiveness	Cardiology	Database of patients receiving intracoronary stents at Duke Heart Center, NC, USA (January 2000–July 2005)	Clopidogrel use in DES and BMS patients	No clopidogrel use in DES and BMS patients	Death, nonfatal MI or a composite of death or MI
O'Donoghue *et al.* (2009)	Comparative effectiveness	Cardiology	The data came from two RCTs: PRINCIPLE–TIMI 44 [53] and TRITON–TIMI 38 [54]	Prasugrel- versus clopidogrel-receiving patients with PPIs	Prasugrel- and clopidogrel-receiving patients without PPIs	Inhibition of platelet aggregation (PRINCIPLE–TIMI 44) Composite of CV death, MI or stroke (TRITON–TIMI 38)

BMS: Bare-metal stent; CER: Comparative effectiveness research; CV: Cardiovascular; DES: Drug-eluting stent; GEE: Generalized estimating equation; HR: Hazard ratio; MI: Myocardial infarction; MIRP: Minimally invasive radical prostatectomy; PPI: Proton pump inhibitor; PS: Propensity score; RCT: Randomized controlled trial; RRP: Retropubic radical prostatectomy.

Table 1. Summary of the selected studies reviewed in this article (cont.).

Estimated treatment effect	Specific PS method used	Downstream impact	Ref.
The estimated HR for each of the four time periods were: ≤180 days: HR: 1.37 (95% CI: 1.27–1.49) <40 days: HR: 1.56 (95% CI: 1.37–1.78) 40–79 days: HR: 1.37 (95% CI: 1.19–1.59) 80–180 days: HR: 1.27 (95% CI: 1.14–1.41)	Cox proportional hazard model was used In confirmatory analysis, Cox models were used in the deciles of the PS [19], which did not change the results based on conventional Cox models	The backdrop of the paper was the US FDA advisory on the potential risks of mortality associated with atypical antipsychotic medications [89]. However, the mortality risk associated with conventional antipsychotics was not known at that time. This paper provided the early evidence of the risks associated with conventional antipsychotic medications. In 2008, 3 years after the publication of this paper, the FDA announced similar warnings for conventional antipsychotics as well [105]	[40]
Between 6 months and 3 years: adjusted relative risk for death for DES compared with BMS cohort was 1.32 (95% CI: 1.11–1.57) At 3 years, adjusted relative risk of death was 1.18 (95% CI: 1.04–1.35)	Cox regression adjustment with estimated PS included as a covariate [19]	The study suggested large RCTs were required to confirm the safety outcomes between DES and BMS cohorts. Recently accumulated RCT evidence suggests comparable safety and effectiveness profiles between DESs and BMSs [49]	[41]
DES cohort at 24 months, which was event free for the first 6 months: death rate: 2.0 vs 5.3% (difference: -3.3%; 95% CI: -6.3 to -0.3%; p = 0.03); death or MI: 3.1 vs 7.2% (difference: -4.1%; 95% CI: -7.6 to -0.6%; p = 0.02); no difference for BMS cohort A similar beneficial effect of clopidogrel continued for the DES cohort that was event free at 12 months, but no difference was observed for the parallel BMS cohort	Inverse weighting with PS and Cox proportional hazard regression was carried out [19,51]	Clopidogrel use appeared to benefit the patients undergoing DES treatment; however, the duration of its use needs to be confirmed through a large RCT	[42]
Mean inhibition of platelet aggregation: clopidogrel cohort: 23.2 ± 19.5% vs 35.2 ± 20.9% (p = 0.02); prasugrel cohort: 69.6 ± 13.5% vs 76.7 ± 12.4% (p = 0.054); primary end point: clopidogrel: HR: 0.94 (95% CI: 0.80–1.11) Prasugrel: HR: 1.00 (95% CI: 0.84–1.20)	PS stratification with Cox regression was fitted within each stratum and the final treatment effect was estimated as the weighted stratum mean	PPIs do not appear to attenuate the antiplatelet effect of clopidogrel. This study demonstrated how clinical trial data may be used to answer a CER question given that the treatment strategy under evaluation was not randomized, and it also highlighted the potential for future consideration of such data for CER research	[43]

BMS: Bare-metal stent; CER: Comparative effectiveness research; CV: Cardiovascular; DES: Drug-eluting stent; GEE: Generalized estimating equation; HR: Hazard ratio; MI: Myocardial infarction; MIRP: Minimally invasive radical prostatectomy; PPI: Proton pump inhibitor; PS: Propensity score; RCT: Randomized controlled trial; RRP: Retropubic radical prostatectomy.

Table 1. Summary of the selected studies reviewed in this article (cont.).

Study (year)	Comparative effectiveness or safety study	Medical specialty	Data source	Treatment	Comparator	Outcome
Hu *et al.* (2009)	Comparative effectiveness	Urology/ general prostate disease	US Surveillance, Epidemiology and End Results (SEER)–Medicare-linked data from 2003 through to 2007	MIRP	Open RRP	The following postoperative outcomes were analyzed: 30-day complications, anastomotic stricture at 31–365 days, incontinence, erectile dysfunction and use of additional cancer therapies
Ray *et al.* (2012)	Comparative safety	Cardiology	Tennessee Medicaid Program Data (1992–2006)	Azithromycin	No antibiotics, amoxicillin, ciprofloxacin or levofloxacin	CV death or all-cause death

BMS: Bare-metal stent; CER: Comparative effectiveness research; CV: Cardiovascular; DES: Drug-eluting stent; GEE: Generalized estimating equation; HR: Hazard ratio; MI: Myocardial infarction; MIRP: Minimally invasive radical prostatectomy; PPI: Proton pump inhibitor; PS: Propensity score; RCT: Randomized controlled trial; RRP: Retropubic radical prostatectomy.

question, given that the treatment strategy under evaluation was not randomized, and it highlighted the potential for the future consideration of such data for CER research.

■ Study 5: Hu *et al.* (2009)

The study from Hu *et al.*, which assessed outcomes following minimally invasive radical prostatectomy (MIRP) versus open retropubic radical prostatectomy (RRP), was a case in point that illustrates the necessity of relying on observational CER due to lack of RCTs [44]. The authors used weighted PS methods to balance observed patient characteristics followed by generalized estimating equation modeling in order to account for potential surgeon clustering [22,55–57]. The primary finding of the study was that, compared with RRP, MIRP was associated with a shorter length of inpatient stay and fewer respiratory and miscellaneous surgical complications and strictures, but was also associated with higher rates of genitourinary complications, incontinence and erectile dysfunction (Table 1) [44].

Table 1. Summary of the selected studies reviewed in this article (cont.).

Estimated treatment effect	Specific PS method used	Downstream impact	Ref.

Complications:
respiratory complications: 4.3 vs 6.6% (p = 0.004);
miscellaneous surgical complications: 4.3 vs 5.6% (p = 0.03);
genitourinary complications: 4.7 vs 2.1% (p = 0.001);
anastomotic stricture: 5.8 vs 14.0% (p < 0.001);
diagnoses of incontinence: 15.9 vs 12.2 per 100 person-years (p = 0.02);
erectile dysfunction: 26.8 vs 19.2 per 100 person-years (p = 0.009);
additional cancer therapies did not differ by surgical procedure: 8.2 vs 6.9 per 100 person-years (p = 0.35)
	Inverse PS weighting [55] followed by GEE use to account for surgeon clustering [57]	Given that there is no head-to-head trial in this area, an observational study such as this was potentially the second-best option. The study clearly shows that while MIRP has some advantages with regards to some of the complications, it is not better than RRP in many other respects, including genitourinary complications, incontinence and erectile dysfunction. Thus, contrary to the well-publicized perception that MIRP is associated with lower postoperative complications, this study documents that patients undergoing MIRP were more likely to have genitourinary complications, incontinence and erectile dysfunction than their RRP counterparts	[44]

Azithromycin vs no antibiotics: CV death: HR: 2.88 (95% CI: 1.79–4.63; p < 0.001); all-cause death: HR: 1.85 (95% CI: 1.25–2.75; p = 0.002); azithromycin vs amoxicillin: CV death: HR: 2.49 (95% CI: 1.38–4.50; p = 0.002) all-cause death: HR: 2.02 (95% CI: 1.24–3.30; p = 0.005)	Pairwise comparison of the treated and control cohorts conducted using PS methods For the azithromycin vs no antibiotic cohort, PS matching was conducted For the azithromycin vs amoxicillin comparison, the distribution of outcomes for amoxicillin was weighted by inverse PS weighting in order to standardize the distribution to that of azithromycin [31] Overlap of the distribution of PS was checked between the intervention and the control cohort. In sensitivity analyses, the results were confirmed through stratified analyses by PS deciles Cox regression models were applied to estimate the HRs	Resulted in safety warning from the FDA on the use of azithromycin [106]	[45]

BMS: Bare-metal stent; CER: Comparative effectiveness research; CV: Cardiovascular; DES: Drug-eluting stent; GEE: Generalized estimating equation; HR: Hazard ratio; MI: Myocardial infarction; MIRP: Minimally invasive radical prostatectomy; PPI: Proton pump inhibitor; PS: Propensity score; RCT: Randomized controlled trial; RRP: Retropubic radical prostatectomy.

This study highlighted the notion that, contrary to the well-publicized perception of MIRP being a complication-free approach, patients undergoing MIRP were more likely to have genitourinary complications, incontinence and erectile dysfunction than their RRP counterparts.

■ Study 6: Ray *et al.* (2012)

Using Tennessee Medicaid program data, Ray *et al.* studied the association between azithromycin use and death (cardiovascular and all-cause) compared with no antibiotics, amoxicillin, ciprofloxacin and levofloxacin [45]. The unit of their analysis was the course of the antibiotic therapy; each azithromycin prescription was frequency-matched to four controls using PS based on 153 covariates [45,58]. Their PS-based analyses demonstrated that azithromycin was associated with an increased risk of cardiovascular and all-cause death compared with the no antibiotics cohort and the amoxicillin cohort (see Table 1 for the specific PS methods used) [45]. Following their publication of these results, the FDA issued a safety warning that azithromycin may be associated with a risk of potentially fatal heart rhythms compared with no drug use, amoxicillin or levofloxacin, and consequently the drug labels were updated in order to reflect this increased risk [45,106].

This study is an excellent example of a PS-based observational CER study that had direct policy implications and potential implications for clinical practice in terms of how azithromycin may be prescribed compared with amoxicillin or levofloxacin.

Discussion

This article has found evidence of a growing use of PS-based observational CER studies over the past decade in five high-impact general and internal medicine journals. One explanation for the expanding use of PS methods in the medical literature is its intuitive appeal, particularly its ability to balance comparison groups on the basis of observed covariates, which seemingly mimics the design of a randomized trial in an observational study [16,19,21,56,59]. PS methods also help to make study analyses more transparent (and seemingly simple), often leading to the presentation of results in a manner that is similar to randomized trials. However, it is worth noting that a detailed assessment of whether PS methods have been applied correctly was beyond scope of this review, and as such, details on the measured covariates adjusted through PS methods, and the potential effect of residual bias due to unmeasured covariates, were not discussed. Our review of the highest-cited articles in high-impact journals over the past decade also highlights the potential for PS-based observational CER studies to have important clinical, policy and research implications. What is noteworthy was that the identified studies were equal in terms of the numbers of comparative safety and comparative effectiveness studies.

There are two clear recommendations for the future observational CER studies using PS methods that emerge from our review. As already noted, PS methods comprise of a collection of different methods [16,19–21,29]. Different PS methods may yield different treatment effect estimates [15,60], potentially leading to different inferences or policy implications. This is why it is important to conduct sensitivity analyses and provide a range of the estimated treatment effects across different methods [15,20,60,61]. For example, PS matching with or without replacement may generate different treatment effect estimates, and so does radius matching with different radii [20]. However, given that different PS methods are not nested, a formal comparison of whether the treatment effect estimates are statistically different or not may not be straightforward, and might require resorting to bootstrap simulations [62]. In a series of recent articles using Monte Carlo simulations, Austin and colleagues have shown how different PS methods perform relatively better than others for estimating specific measures, such as marginal odds ratios, relative risks and hazard ratios [60,63–68]. These papers, however, do not provide general guidance to a CER practitioner in choosing the appropriate PS method for the specific data situation in hand. Nevertheless, in order to ensure that the CER results are robust to the specific PS methods used, simple sensitivity analyses may comprise of tabulating the treatment effect estimates from different PS methods, which would then provide a range for the treatment effects for different PS methods. Bootstrap methods can then be used to assess whether the estimates are different from each other.

Previous systematic reviews suggest that the implementation and reporting of PS methods in clinical studies have not been well documented [14,17,32]. Our review of the six studies in this article also bears testimony to the above two issues. Even these highly cited studies published in high-impact journals, which is presumably a proxy for the quality of the papers, used a specific PS method (e.g., matching [45], stratification [40,43], covariate adjustment [41] and inverse probability weighting [42,44]); however, it is unclear whether and how the reported results would have changed if those studies had used alternative PS methods. Another important aspect of PS implementation is assessing the extent to which comparative effectiveness results derived from PS methods such as matching are robust to the presence of potential unmeasured confounders through sensitivity analyses [69–71]. However, this important aspect of PS implementation has been largely ignored in the literature, including the six papers reviewed in our study. This may have been due to a lack of available software in order to conduct such sensitivity analyses, which has been addressed in recent times [72,107].

In addition, the studies included in our review reported various aspects of PS implementation, including checking of overlap of the PS distribution between treated and control patients and covariate balance testing in a nonuniform manner. Thus, the field of CER would benefit from a uniform set of implementation and reporting guidelines for PS methods. Such guidelines need to be developed collaboratively by different stakeholders (e.g., industry, academia, payers and government), which will improve transparency and the widespread adoption of PS methods. Peer-reviewed scientific journals can also contribute to the uniform adoption of PS guidelines by requiring authors of CER studies to document the use of these guidelines. Several organizations have either developed or are currently working to help develop guidelines for observational research. For example, a recent report by the Agency for Healthcare Research and Quality (AHRQ) provides guidance on the methodological issues surrounding observational CER [108]. In addition, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) are both working to develop standards for observational research [109,110]. This approach can also help to minimize the problem of selective reporting (both of the outcomes and the specific methods used) that often characterize observational studies [73]. In addition, the proposed guidelines may require all observational CER studies to be registered, so that all of the relevant issues in conducting a specific observational study, including the statistical methods to be used for evaluating comparative effectiveness, are prespecified, which in turn will address the issue of selective reporting and also improve the credibility of the results [6,7,73]. The ultimate goal of mandating guideline adoption for CER studies using PS is to ensure that the reported evidence of CER is robust and reproducible.

The second recommendation relates to alleviating the inherent issue that due to PS methods' inability to adjust for unmeasured characteristics. In other words, PS methods are not panacea for the lack of crucial data on treatment assignment, including patient and provider characteristics, access to healthcare and all of the relevant clinical information that might influence the specific intervention that the patient receives instead of the alternatives. In order for observational studies to become more reliable, these data gaps need to be bridged. For example, observational databases need to capture more clinical data [74], which can potentially be accomplished by merging claims databases with clinical databases (e.g., laboratory databases or patient/disease registries). Of course, the completeness of the data fields and/or minimal loss to follow-up of patients in such clinical databases is going to be key to ensuring the validity of their use in CER research. Demographic and other characteristics can influence treatment choice in addition to the outcomes of the treatment [1], as well as patient preferences, which, in turn, can potentially impact provider decisions on the choice of specific interventions [75]. Wherever possible, patient centeredness of the research design and the corresponding outcomes may be improved by linking observational databases to the qualitative data gathered from focus group or in-depth interviews. Thus, capturing these patient characteristics and/or merging multiple databases that facilitate access to a rich set of confounders will minimize the unmeasured bias associated with PS methods. The good news is that there has been a movement towards achieving this goal (e.g., through the recent initiatives for building distributed data networks, with or without a central secure data warehouse; the former houses multiple data sources, including electronic health records [76,77], while in the latter arrangement, data are distributed across the network and often behind the firewall, including the Surveillance, Epidemiology and End Results [SEER]–Medicare linked data [111] and the Society for Thoracic Surgeons [STS] national database linked to Medicare claims data [78]). The importance of developing a national infrastructure for facilitating robust CER, which would also address many of the issues that we have discussed, was echoed in two recent Institute of Medicine (IOM) reports [79,80]. One hurdle towards achieving the goal of merging multiple datasets is patient privacy, which might be compromised. But there are ways to overcome these privacy concerns, as shown by Rassen *et al.*, which enabled data pooling from multiple sources followed by multivariate adjustment on the fly, without requiring patient-identifying information [81].

Several limitations of the current review should be acknowledged. Although the formal CER definition covers a comprehensive list of interventions that includes any strategy or item "used in the treatment, management, diagnosis of or prevention of illness or injury in individuals" [101], we restricted our review only to those studies that involved a comparison of non-placebo pharmaceutical drugs, medical/surgical procedures or medical devices for tractability. Furthermore, our literature search results on trends in PS-based observational CER studies may not be generalizable, given our selection of only high-impact journals in the field of general and internal medicine. We used this selection strategy and the number of citations as a proxy for the quality and potential impact of articles to be identified for the review. However, this clearly may not be the case, given the subjectivity of the peer-review process across all journals. Additionally, we restricted the full-text review to the top five cited articles in each 2-year timeframe as a convenient sample. The choice of the top five instead of another number was, admittedly, arbitrary. Had we allowed all articles found from the methodology of searching journals to undergo a full-text review, we may have had more articles satisfy the study inclusion criteria. Finally, a large literature – particularly in health econometrics – has attempted to address the bias issues introduced by unobserved confounders through alternative methodologies, such as instrumental variables (IVs) [82,83]. While a full discussion of these methods is beyond the scope of this review, it should be noted that these methods bear many similarities to PS-based methods. In fact, the estimation of PS is often a preliminary step in the application of these related tools, including the local average treatment effect estimator, which is a ratio of two matching estimators with the denominator indicating the difference in propensity of receiving treatment at two values of the IV [69,84]. The key distinguishing feature of IV methods is that they can address both observed and unobserved confounders, whereas PS analysis, as discussed previously, is limited in its ability to account for the latter. Nevertheless, the application of IV methods in CER is largely limited due to the difficulty in identifying reliable and valid instruments [85]. Furthermore, as opposed to the simplicity of the PS methods, the interpretation of the treatment effects in IV methods are not intuitive to most lay audiences. Regardless, discordance in the estimates of treatment effects may arise between PS and IV methods [86,87], which may be confusing for the end user of the research findings, suggesting that CER evidence from observational data needs to be validated with alternative methods [40,86,87].

Conclusion

Although there are numerous well-known challenges to the robustness of conclusions drawn from any observational study, we believe that the combination of improving data systems and statistical methodology, including the ability to determine the most appropriate PS method to use in a specific data situation, will play an important role in helping to address these challenges. PS methods will continue to remain a key part of the statistical toolkit for researchers conducting CER studies using observational data.

Future perspective

It seems clear that the demand for evidence from observational studies will continue to grow in the coming years. The need for evidence in order to provide guidance to clinicians, policymakers and patients is acute. There are simply too many questions in healthcare that it is not feasible to answer them all with RCTs, nor can society afford the time and resources associated with developing all CER evidence through randomized trials. Although there are numerous well-known challenges to the robustness of the conclusions drawn from any observational study, we believe that the combination of improving data systems and statistical methodologies will play an important role in helping to address these challenges. This involves not only the linkage of claims and electronic health record data, but also patient-centered outcomes and qualitative data. PS methods will be a key part of the statistical toolkit for researchers conducting CER using observational data.

Acknowledgements

The authors would like to sincerely acknowledge the assistance of ML Rethlefsen of the Mayo Clinic Libraries in searching for the relevant literature citations used in this manuscript.

Financial & competing interests disclosure

BJ Borah and JP Moriarty were partially funded by an internal award from the Division of Health Care Policy and Research, Mayo Clinic (MN, USA). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

- While randomized controlled trials (RCTs) are considered to have the strongest research design, it is not possible to conduct an RCT to answer every comparative effectiveness research (CER) question. Moreover, RCTs are generally time consuming and expensive to conduct relative to observational studies, and they often lack generalizability to real-world settings. Given the growing demand for CER to inform real-world clinical practice, there will be a growing demand for evidence from observational studies.
- Given the challenges to drawing reliable conclusions from observational studies, it is very important to use appropriate research designs and statistical methods in order to account for the biases in observational data.
- Propensity score (PS) methods are an important set of tools for the analysis of observational data, and their use has been growing substantially in recent years.

Literature search & study selection strategy

- Potential studies were identified in Scopus, a multidisciplinary citation index covering 20,000 peer-reviewed journals in biomedicine and health, science, technology and the humanities from 1995 to the present.
- The search was restricted to the top five (in terms of impact factor) general and internal medicine journals publishing original research.
- A full-text review was performed on the six articles with the highest impact factor among each of the 2-year timeframes spanning from 2001 to 2012. These studies used PS methods in order to adjust for baseline confounds, while evaluating the comparative effectiveness of the interventions under study.

Overview of selected studies

- Study 1 looked at the comparable risk profile of both typical and atypical antipsychotic medications. The findings of this study had implications for the generation of a US FDA advisory.
- Study 2 evaluated the hypothesis that drug-eluting stents (DESs) were associated with a higher rate of long-term adverse outcomes compared with bare-metal stents, and found evidence to support this. This study highlighted the need for a large RCT in order to further validate the findings of higher adverse outcomes in DESs.
- Study 3 investigated whether a shorter regimen of clopidogrel therapy was associated with a higher incidence of deaths and/or myocardial infarction in percutaneous coronary intervention patients with DESs and bare-metal stents. The results highlighted the notion that clopidogrel use appeared to have benefited patients undergoing DES treatment, although the duration of its use needs to be confirmed through a large RCT.
- Study 4 found that concomitant use of proton pump inhibitors was not suggestive of an attenuation of the effects of clopidogrel versus prasugrel. This study was a novel application of data collected in clinical trials in order to answer a CER question for which the specific intervention being evaluated was not randomized, and it highlighted the potential for future consideration of such data for CER.
- Study 5 assessed outcomes following minimally invasive radical prostatectomy versus open retropubic radical prostatectomy. This study highlighted the fact that, contrary to the well-publicized perception that minimally invasive radical prostatectomy is a complication-free approach, patients undergoing this procedure were at higher risk of developing genitourinary complications, incontinence and erectile dysfunction than their retropubic radical prostatectomy counterparts.
- Study 6 studied the association between azithromycin use and death compared with no antibiotics, amoxicillin, ciprofloxacin and levofloxacin. Following the publication of this study, the FDA issued a safety warning that azithromycin may be associated with a risk of potentially fatal heart rhythms compared with no drug use, amoxicillin or levofloxacin, and consequently the drug labels were updated to reflect this increased risk.

Discussion

- The current article has found a growing use of PS-based observational CER studies over the past decade in high-impact general and internal medicine journals.
- The field of CER requires a uniform set of implementation and reporting guidelines for PS methods. Such guidelines need to be developed collaboratively by different stakeholders (e.g., industry, academia, payers and government), which will improve transparency and the widespread adoption of PS methods.
- PS methods are not a panacea for a lack of crucial data on treatment assignment, including patient and provider characteristics, access to healthcare and all of the relevant clinical information that might influence the specific intervention that the patient receives instead of the alternatives. In order for observational studies to become more reliable, these gaps in the data need to be bridged. Furthermore, the use of PS methods is not a panacea for improper study design.
- Although there are numerous well-known challenges to the robustness of the conclusions drawn from any observational study, we believe that the combination of improving data systems and statistical methodologies, such as PS methods discussed in the current article, will play an important role in helping to address these challenges.

References

Papers of special note have been highlighted as:

■ of interest

■■ of considerable interest

1 Garber AM, Sox HC. The role of costs in comparative effectiveness research. *Health Aff. (Milwood)* 29(10), 1805–1811 (2010).

2 D'Agostino RB Jr, D'Agostino RB Sr. Estimating treatment effects using observational data. *JAMA* 297(3), 314–316 (2007).

3 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J. Chronic Dis.* 20(8), 637–648 (1967).

4 Tunis SR, Stryer DB, Clancy CM. Practical clinical trials – increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 290(12), 1624–1632 (2003).

5 Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet* 363(9422), 1728–1731 (2004).

6 Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *CMAJ* 182(15), 1638–1642 (2010).

7 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.* 33(15), 1893–1901 (2012).

■ **Finds that divergent results generated in randomized controlled trials (RCTs) and observational studies using propensity score (PS) methods as they apply to acute coronary syndrome were not statistically significant.**

8 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med.* 342(25), 1878–1886 (2000).

9 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* 342(25), 1887–1892 (2000).

10 Kuss O, Legler T, Bogermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *J. Clin. Epidemiol.* 64(10), 1076–1084 (2011).

■ **Compares treatment effects between RCTs and PS analyses of coronary artery bypass grafting, showing that the treatment effects from RCTs and PS analyses were very similar.**

11 McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 319(7205), 312–315 (1999).

12 Albanes D. Beta-carotene and lung cancer: a case study. *Am. J. Clin. Nutr.* 69(6), 1345S-1350S (1999).

13 Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. *N. Engl. J. Med.* 348(7), 645–650 (2003).

14 Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat. Med.* 27(12), 2037–2049 (2008).

15 Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat. Med.* 25(12), 2084–2106 (2006).

16 Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin. Pharmacol. Toxicol.* 98(3), 253–259 (2006).

17 Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J. Clin. Epidemiol.* 58(6), 550–559 (2005).

18 Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J. Clin. Epidemiol.* 59(5), 437–447 (2006).

19 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).

20 Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econom. Stat.* 84(1), 151–161 (2002).

21 D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* 17(19), 2265–2281 (1998).

22 Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79(387), 516–524 (1984).

23 Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Stat. Assoc.* 95(450), 573–585 (2000).

24 Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 11(5), 561–570 (2000).

25 Rosenbaum PR. Model-based direct adjustment. *J. Am. Stat. Assoc.* 82(398), 387–394 (1987).

26 Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4), 962–972 (2005).

27 Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20(4), 512–522 (2009).

28 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28(25), 3083–3107 (2009).

29 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46(3), 399–424 (2011).

30 Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am. J. Epidemiol.* 163(12), 1149–1156 (2006).

31 Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am. J. Epidemiol.* 161(9), 891–898 (2005).

32 Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol. Drug Saf.* 13(12), 841–853 (2004).

33 Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann. Intern. Med.* 137(8), 693–695 (2002).

34 Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* 158(3), 280–287 (2003).

35 Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 49(4), 1231–1236 (1993).

36 Rubin DB. Estimating the causal effects of smoking. *Stat. Med.* 20(9–10), 1395–1414 (2001).

37 Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* 26(1), 20–36 (2007).

38 Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am. J. Epidemiol.* 162(3), 279–289 (2005).

39 Eng PM, Seeger JD, Loughlin J, Clifford CR, Mentor S, Walker AM. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiol. Drug Saf.* 17(3), 297–305 (2008).

40 Wang PS, Schneeweiss S, Avorn J *et al.* Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *N. Engl. J. Med.* 353(22), 2335–2341 (2005).

■■ **Study using PS methods to compare patients using conventional antipsychotic medications with those using atypical antipsychotic medications, finding that they had similar risks.**

41 Lagerqvist B, James SK, Stenestrand U *et al.* Long-term outcomes with drug-eluting stents versus bare-metal stents in Sweden. *N. Engl. J. Med.* 356(10), 1009–1019 (2007).

■■ **Study using PS as a covariate in a regression analysis to show that drug-eluting stents were associated with a higher risk of long-term adverse events compared with bare-metal stents.**

42 Eisenstein EL, Anstrom KJ, Kong DF *et al.* Clopidogrel use and long-term clinical outcomes after drug-eluting stent implantation. *JAMA* 297(2), 159–168 (2007).

■■ **Study using inverse PS methods in a regression analysis to show that extended clopidogrel use in drug-eluting stent patients was associated with a reduced risk of adverse events.**

43 O'Donoghue ML, Braunwald E, Antman EM *et al.* Pharmacodynamic effect and clinical efficacy of clopidogrel and prasugrel with or without a proton-pump inhibitor: an analysis of two randomised trials. *Lancet* 374(9694), 989–997 (2009).

■■ **Study using PS methods to find that concomitant use of proton pump inhibitors was not suggestive of an attenuation of the effects of clopidogrel versus prasugrel use.**

44 Hu JC, Gu X, Lipsitz SR *et al.* Comparative effectiveness of minimally invasive vs open radical prostatectomy. *JAMA* 302(14), 1557–1564 (2009).

■■ **Study using weighted PS methods and the generalized estimating equations to find that minimally invasive radical prostatectomy was associated with better outcomes than retropubic radical prostatectomy.**

45 Ray WA, Murray KT, Hall K, Arbogast PG, Stein CM. Azithromycin and the risk of cardiovascular death. *N. Engl. J. Med.* 366(20), 1881–1890 (2012).

■■ **Study using a PS-matched cohort demonstrating that azithromycin was associated with an increased risk of adverse events compared with no antibiotics and amoxicillin.**

46 Gill SS, Bronskill SE, Normand SL *et al.* Antipsychotic drug use and mortality in older adults with dementia. *Ann. Intern. Med.* 146(11), 775–786 (2007).

47 Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. *CMAJ* 176(5), 627–632 (2007).

48 Stefanini GG, Holmes DR Jr. Drug-eluting coronary-artery stents. *N. Engl. J. Med.* 368(3), 254–265 (2013).

49 Ong AT, Hoye A, Aoki J *et al.* Thirty-day incidence and six-month clinical outcome of thrombotic stent occlusion after bare-metal, sirolimus, or paclitaxel stent implantation. *J. Am. Coll. Cardiol.* 45(6), 947–953 (2005).

50 Pfisterer M, Brunner-La Rocca HP, Buser PT *et al.* Late clinical events after clopidogrel discontinuation may limit the benefit of drug-eluting stents: an observational study of drug-eluting versus bare-metal stents. *J. Am. Coll. Cardiol.* 48(12), 2584–2591 (2006).

51 Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics* 57(4), 1207–1218 (2001).

52 Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 87(2), 329–343 (2000).

53 Wiviott SD, Trenk D, Frelinger AL *et al.*; PRINCIPLE-TIMI 44 investigators. Prasugrel compared with high loading- and maintenance-dose clopidogrel in patients with planned percutaneous coronary intervention: the Prasugrel in Comparison to Clopidogrel for Inhibition of Platelet Activation and Aggregation–Thrombolysis in Myocardial Infarction 44 trial. *Circulation* 116(25), 2923–2932 (2007).

54 Wiviott SD, Braunwald E, McCabe CH *et al.* Prasugrel versus clopidogrel in patients with acute coronary syndromes. *N. Engl. J. Med.* 357(20), 2001–2015 (2007).

55 Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5), 550–560 (2000).

56 Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127(8 Pt 2), 757–763 (1997).

57 Zeger SL, Liang KY. Longitudinal data-analysis for discrete and continuous outcomes. *Biometrics* 42(1), 121–130 (1986).

58 Rothman KJ, Greenland S. *Modern Epidemiology (2nd Edition).* Lippincott-Raven, PA, USA (1998).

59 Rosenbaum PR, Rubin DB. Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statistician* 39(1), 33–38 (1985).

60 Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat. Med.* 26(4), 734–753 (2007).

61 Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value Health* 9(6), 377–385 (2006).

62 Efron B, Tibshirani RJ. *An Introduction to Bootstrap*. Chapman and Hall, NY, USA (1993).

63 Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat. Med.* 26(16), 3078–3094 (2007).

64 Austin PC. The performance of different propensity-score methods for estimating relative risks. *J. Clin. Epidemiol.* 61(6), 537–545 (2008).

65 Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med. Decis. Making* 29(6), 661–677 (2009).

66 Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom. J.* 51(1), 171–184 (2009).

67 Austin PC. Different measures of treatment effect for different research questions. *J. Clin. Epidemiol.* 63(1), 9–10 (2010).

68 Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat. Med.* 32(16), 2837–2849 (2013).

69 Diprete TA, Gangl M. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociol. Methodol.* 34(1), 271–310 (2004).

70 Rosenbaum PR. *Observational Studies (2nd Edition).* Springer, NY, USA (2002).

71 Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* 15(5), 291–303 (2006).

72 Becker SO, Caliendo M. Sensitivity analysis for average treatment effects. *Stata J.* 7(1), 71–83 (2007).

73 Collins GS, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting. *Eur. Heart J.* 33(15), 1867–1869 (2012).

74 Sox HC. Defining comparative effectiveness research: the importance of getting it right. *Med. Care* 48(6 Suppl.), S7–S8 (2010).

75 Barry MJ, Mulley AG Jr, Fowler FJ, Wennberg JW. Watchful waiting vs immediate transurethral resection for symptomatic prostatism. The importance of patients' preferences. *JAMA* 259(20), 3010–3017 (1988).

76 Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med. Care* 48(6 Suppl.), S45–S51 (2010).

77 Libby AM, Pace W, Bryan C *et al.* Comparative effectiveness research in DARTNet primary care practices: point of care data collection on hypoglycemia and over-the-counter and herbal use among patients diagnosed with diabetes. *Med. Care* 48(6 Suppl.), S39–S44 (2010).

78 Jacobs JP, Edwards FH, Shahian DM *et al.* Successful linking of the Society of Thoracic Surgeons adult cardiac surgery database to Centers for Medicare and Medicaid Services Medicare data. *Ann. Thorac. Surg.* 90(4), 1150–1156; discussion 1156–1157 (2010).

79 Institute of Medicine (IOM). *Learning What Works: Observational Studies in a Learning Health System: Workshop Summary.* The National Academies Press, Washington, DC, USA (2013).

▪ **Summary of an Institute of Medicine workshop discussing the methodologies needed to fill gaps in comparative effectiveness research in the USA.**

80 Institute of Medicine (IOM). *Learning What Works: Infrastructure Required for Comparative Effectiveness Research.* The National Academies Press, Washington, DC, USA (2011).

81 Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med. Care* 48(6 Suppl.), S83–S89 (2010).

82 Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol. Drug Saf.* 19(6), 537–554 (2010).

83 Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Health Econ.* 27(3), 531–543 (2008).

84 Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91(434), 444–455 (1996).

85 Davies NM, Smith GD, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology* 24(3), 363–369 (2013).

86 Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 297(3), 278–285 (2007).

87 Venkitachalam L, Lei Y, Magnuson EA *et al.* Survival benefit with drug-eluting stents in observational studies: fact or artifact? *Circ. Cardiovasc. Qual. Outcomes* 4(6), 587–594 (2011).

88 Kuehn BM. FDA warns antipsychotic drugs may be risky for elderly. *JAMA* 293(20), 2462 (2005).

▪ Websites

101 US Government. The Patient Protection and Affordable Care Act, Public Law 111–148, 111th Congress (2010). www.gpo.gov/fdsys/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf

102 PubMed homepage. www.ncbi.nlm.nih.gov/pubmed

103 Scopus homepage. www.scopus.com

104 Thomson Reuters. Web of Science – Journal Citation Reports. http://wokinfo.com/products_tools/analytical/jcr

105 US FDA. Information for healthcare professionals: conventional antipsychotics (2008). www.fda.gov/drugs/afetyafetyinformationforpatientsandproviders/ucm124830.htm

106 US FDA. FDA drug safety communication: Azithromycin (Zithromax or Zmax) and the risk of potentially fatal heart rhythms (2013). www.fda.gov/drugs/drugsafety/ucm341822.htm

107 Keele L. An overview of rbounds: an R package for Rosenbaum bounds sensitivity analysis with matched data (2010). www.personal.psu.edu/ljk20/rbounds%20vignette.pdf

108 Developing a Protocol for Observational Comparative Effectiveness Research. www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf

109 International Society for Pharmacoeconomics and Outcomes Research. Outcomes research guidelines index. www.ispor.org/guidelinesindex

110 European Network of Centres for Pharmacoepidemiology and Pharmacovigilance Standards and Guidances. www.encepp.eu/standards_and_guidances/index.shtml

111 National Cancer Institute (NCI). SEER–Medicare Linked Database. http://appliedresearch.cancer.gov/seermedicare