

Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review

Journal of **Comparative Effectiveness Research**

Lucas Malla^{*,1}, Rafael Perera-Salazar², Emily McFadden², Morris Ogero³, Kasia Stepniewska^{1,4} & Mike English^{1,3}

¹Nuffield Department of Medicine, University of Oxford, Oxford, UK

²Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

³Kenya Medical Research Institute-Wellcome Trust Research Programme, Nairobi, Kenya

⁴WorldWide Antimalarial Resistance Network, Oxford, UK

* Author for correspondence: lmalla@kemri-wellcome.org

Aim: Even though systematic reviews have examined how aspects of propensity score methods are used, none has reviewed how the challenge of missing data is addressed with these methods. This review therefore describes how missing data are addressed with propensity score methods in observational comparative effectiveness studies. **Methods:** Published articles on observational comparative effectiveness studies were extracted from MEDLINE and EMBASE databases. **Results:** Our search yielded 167 eligible articles. Majority of these studies (114; 68%) conducted complete case analysis with only 53 of them stating this in the methods. Only 16 articles reported use of multiple imputation. **Conclusion:** Few researchers use correct methods for handling missing data or reported missing data methodology which may lead to reporting biased findings.

First draft submitted: 13 September 2017; Accepted for publication: 22 September 2017; Published online: 5 October 2017

Keywords: comparative effectiveness • missing data • propensity score

Randomized controlled trials (RCTs) are generally considered the gold standard for estimating causal effects of treatments [1]. If well-designed and executed, randomization ensures that both measured and unmeasured factors are comparable between treatment groups, meaning that any effects seen are likely to be due to the treatment and not other factors [2]. However, RCTs may not always be an ethical or practical option, and researchers are increasingly making use of routinely collected (observational) data in an attempt to answer questions that might be difficult to address using a trial design [3], for example to investigate potential benefits and harms of interventions that are already accepted in general practice, or in different populations, or under different conditions, to give results that are potentially more generalizable than those from RCTs [4]. Two key analytic considerations of these studies are: the methods used to account for differences in patient characteristics, usually in an effort to create groups similar to those that might have arisen after randomization, and the methods to handle missing data. In routine clinical practice, patient allocation to alternative treatments for the same illness may be conditioned on varying characteristics, and thus, any two or more treatment groups may be different. To examine differences in treatment outcome(s), there is therefore a need to make groups comparable or 'enforce' covariate balance prior to comparative analysis [5]. The ideal approach is to compare individuals with exactly similar values on observed covariates. However, this approach may be inefficient when dealing with a larger number of covariates as it may be difficult to find sufficient individuals with exactly similar covariate values in treatment groups [6]. An alternative that is increasingly being used in the last decade (across many disciplines) is the propensity score (PS) approach [7,8] developed by Rosenbaum and Rubin (1983) [9]. Here, the PS describes the probability of a patient's assignment into a treatment group given the observed covariates. Outcomes for patients with similar PSs (between treatment groups) are then comparable as it is assumed they have similar distribution of observed covariates. Estimation of PSs requires that all covariate data are fully observed. However in case of missing data, the patients with missing data are excluded from analysis particularly when using non-Bayesian models that are most commonly applied in practice.

Future
Medicine

Three methods may be used in the face of missing data, these include: complete case analysis, multiple imputation (MI) and estimation of PS within various patterns of missing data [10]. These methods are briefly discussed.

Complete case analysis

This method excludes all patients if they have missing data in at least one of the covariates (or waves of follow-up in a cohort study) from the analysis leading to loss of power. This approach only results in unbiased estimates when it is demonstrable that missing data are unrelated to the study treatments or design. For instance, if a cancer patient is enrolled in a follow-up study and relocates to another city such that subsequent measurements may not be obtained – then such data may be understood to be missing completely at random (MCAR) and estimation of PS and outcome analysis excluding such patients may be valid.

Multiple imputation

MI involves filling in plausible values for missing data more than once, which accounts for uncertainty in filling in unknown values [11]. MI is commonly practiced under the assumption that data are missing at random (MAR) – and this means that the missing data may potentially be related to observed co-variables but not those that are unobserved [12]. For instance, in a follow-up study, if a patient's condition was consistently improving over previous visits and all of a sudden she/he drops out – then it might be plausible to assume the patient is doing better (based on previous information).

As MI generates multiple datasets, two approaches have been proposed to estimate PS – within and across these [13–15]. In the within approach, PSs (per participant) are estimated for each of the multiple imputed datasets. Then the appropriate PS method is used on each of these datasets to create balanced patient groups with treatment effects also estimated within each dataset. While, in the across approach, the estimated multiple PSs per patient are averaged across the imputed datasets – then used to estimate treatment effects in each imputed dataset. Both the within and across approaches result in multiple treatment effects which are pooled using Rubin rules [14] to account for within and between imputation variability.

Propensity score estimation by patterns of missing data

This method aims to retain all the patients in an analysis by grouping them based on observed covariates and using models to estimate PS in each group. If there are clearly defined groupings in the study related to the design – and some data may not be obtained due to this, then clearly such data are missing not at random (MNAR). And this may be a valid method under the MNAR assumption.

Other less intuitive methods have historically been used to address missing data and these include the use of the: previous value to replace missing data (last observation carried forward [LOCF]), substitution with population mean value (simple imputation) and missing indicator method – where those with missing data are assigned to a group. These methods do not account for the uncertainty involved in filling in plausible values [16], and almost always bias overall estimates particularly if they are not aligned with any missingness assumption.

The three missing data assumptions defined (MCAR, MAR and MNAR) are generally not directly testable but are instead justifiable through understanding and describing the process that generated the data.

Existing systematic reviews [7,8,17] have focused on the use and adequacy of reporting of PS methods with an aim of examining reproducibility of findings. The methodological aspects examined in these reviews included: variable selection, methods for estimating and using PS, checking of covariate balance, methods for analysing outcome(s) after PS estimation among others. However, none of the reviews examined how PS methods are used in the presence of missing data. Analysis may result in biased estimates if missing data are not correctly handled [18] even with proper use and reporting of PS methods. On the other hand, existing missing data reviews have focused on RCTs or observational studies that do not focus on comparative evaluations (for example see [19–21]).

Objective

Building on previous systematic reviews [7,8,17,21,22], we therefore systematically reviewed published observational comparative effectiveness studies primarily to assess the methods for handling missing data in estimation of PS and if the methods used are in line with the STROBE guidelines [23]. The STROBE guidelines make recommendations for how missing data should be reported in observational studies and in particular they require researchers to report

on the proportion of missing data, reasons why data are missing and how missing data are addressed. These three elements are examined to determine adherence to STROBE guidelines.

Systematic review methods

Search term & literature search databases

We used the search terms ("*observational stud** and *propensity score**") – and with this literature searches were carried out in Embase(OvidSP) [1974 to 30 June 2017] and Medline(OvidSP) [1946 to June 30, 2017]. The searches were further restricted to clinical articles published in the last 7.5 years (since January 2010). The restriction was motivated by the fact that the STROBE guidelines were first published in 2007 [23] and researchers would actively use them post 2007.

Inclusion & exclusion criteria

Primary research studies of observational comparative effectiveness studies (using actual patient data), published in English, were included. Studies only discussing methods in randomized studies, demonstrating methods using simulated data, meta-analyses and systematic reviews were therefore excluded, although bibliographies from secondary studies were hand searched for additional relevant primary observational studies. Conference abstracts were also excluded due to the limited information they provided. Further, studies on quasi experiments were also excluded since a researcher may sometimes have control over treatment assignment. All the studies excluded as described here are referred to as nonprimary.

Screening of papers & data extraction

Abstracts were selected for text review and final inclusion by one reviewer (L Malla). If any abstract did not offer sufficient information, then the full text was obtained (together with any available corresponding supplementary materials) and scanned through for the relevant aspects of method(s). One reviewer (L Malla) extracted data into structured forms (from the full text of the selected articles) which contained variables specific to: article characteristics; and missing data methods (see [Table 1](#) for full description of the variables). A second reviewer (M Ogero) randomly selected 25% ($n = 42$) of the articles that met inclusion and exclusion criteria and did an independent data abstraction then reviewed methods, and the degree of agreement was examined using Kappa coefficient [24] and is reported. This comparison in agreement was examined for each of: proportion of missing data reported, reasons for missing data indicated and method used to address missing data. Where there was any disagreement, L Malla and M Ogero jointly reviewed the article and reached a consensus.

Presentation of findings

Articles were stratified into four publication time periods; 2010–2011, 2012–2013, 2014–2015 and 2016–2017 to explore trends in practice. Frequencies and proportions were used to summarize findings. This review followed the various methodological aspects recommended in the PRISMA statement [25].

Results

Study characteristics

The process of identification and selection of articles is summarized in [Figure 1](#) (adapted from the PRISMA flow diagram). We identified 2422 articles after 973 duplicates were removed, 2255 articles did not meet inclusion criteria leaving 167 articles for full text review.

71% ($n = 118$) of the eligible articles retrospectively analyzed routine datasets, most of which were derived from medical registry databases. The remaining 29% ($n = 49$) used prospective designs which were based in hospital settings. Most of the articles (90%; $n = 151$) conducted comparative analyses using two treatment groups while 10% ($n = 16$) reported analyses using more than two groups. Almost half of the articles ($n = 81$) were based on data from North America (USA and Canada) – while 29% ($n = 49$), 18% ($n = 30$), 4% ($n = 6$) and 1% ($n = 1$) were based on data from Europe, Asia, Australia and Africa, respectively ([Table 2](#)). See the extracted data and the corresponding references for reviewed articles presented as [Supplementary Appendix 1 & 2](#).

The Kappa estimates of agreement on the selected missing data methodological aspects independently identified by both L Malla and M Ogero ranged from 0.96 to 1 which indicated excellent agreements (see [Supplementary Appendix 3](#) for Kappa calculations).

Table 1. List of variables extracted.		
#	Variable	Response options
Article characteristics		
1	Year of publication	–
2	Title of publication	–
3	Number of treatment groups compared	(1) 2 (2) >2
4	Setting (country/continent)	–
Missing data methods		
5	Proportion of missing data reported	(1) Yes (2) No
6	Missing data method reported	(1) Yes (2) No
7	Missing data mechanism mentioned	(1) Yes (2) No
8	Reason for missing data given	(1) Yes (2) No
9	Specific missing data mechanism	(1) MCAR (2) MAR (3) MNAR (4) Not mentioned
10	Specific missing data method used	(1) Complete case (2) Pattern mixture (3) Multiple imputation (4) Not mentioned
11	Missing data sensitivity conducted	(1) Yes (2) No
12	Analysis compared between those with complete and incomplete data	(1) Yes (2) No
13	Variables included in MI explained (if MI used)	(1) Yes (2) No
14	Number of imputations specified (if MI used)	(1) Yes (2) No
15	Methods used to estimate propensity scores after MI	(1) Within (2) Across (3) Not mentioned
16	Software used for MI	(1) R (Hmisc, MICE, mi, etc.) (2) SAS (MI) (3) STATA (MI) (4) Other (5) Not mentioned
The values in brackets indicate option numbers. MAR: Missing at random; MCAR: Missing completely at random; MI: Multiple imputation; MNAR: Missing not at random.		

Table 2. Description of the studies.	
Study characteristics	n (%)
Design	
Retrospective	118 (71%)
Prospective	49 (29%)
Setting	
North America	81 (49%)
Europe	49 (29%)
Asia	30 (18%)
Australia	6 (4%)
Africa	1 (1%)
Number of treatments compared	
2	151 (90%)
>2	16 (10%)

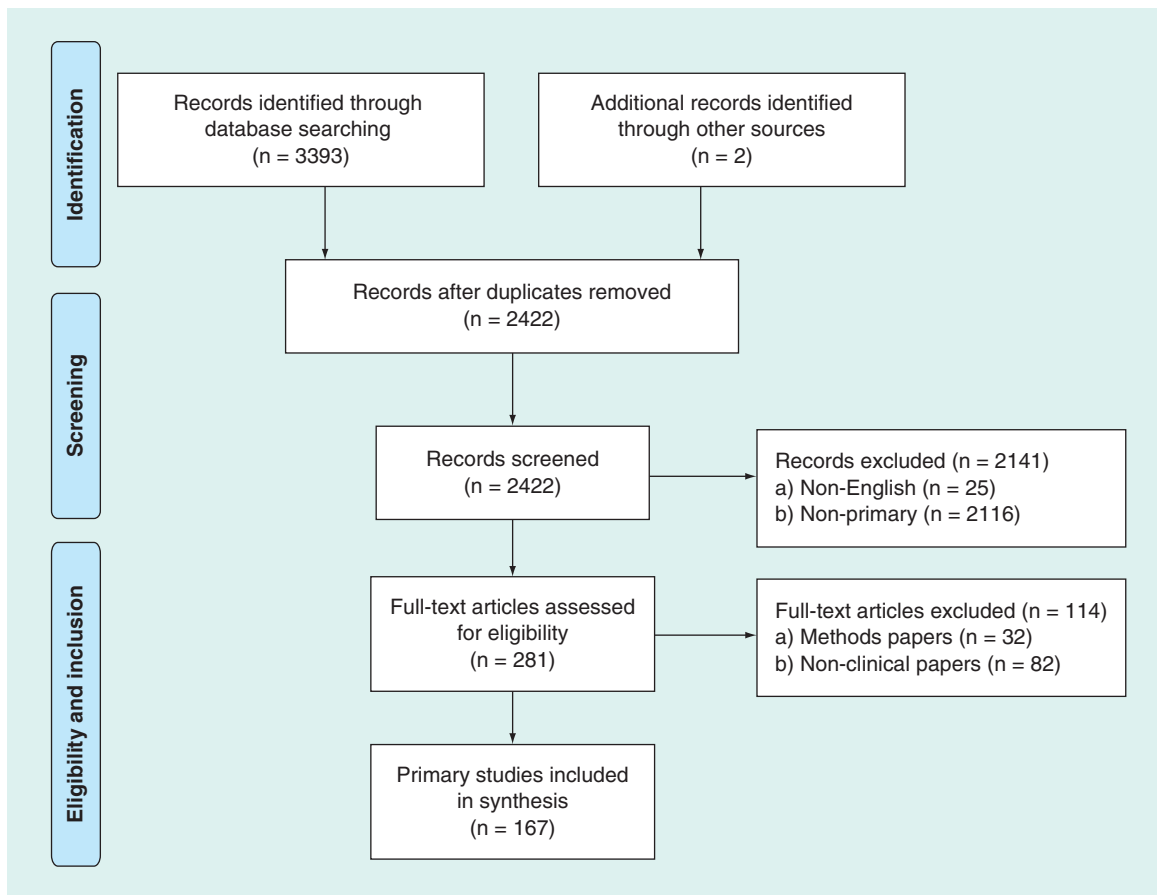


Figure 1. Selection process of primary observational studies.

Reporting of missing data

Of the 167 articles, 37% ($n = 62$) provided information on the amount of missing data and only 12 provided reasons for missingness. Of the 12, three articles linked the reasons to the type of missing data mechanism: MAR ($n = 1$) and MCAR ($n = 2$). None of the remaining articles that provided information on amount of missing data ($n = 59$) commented on any assumed missingness mechanism.

Missing data methods used

Missing data were addressed in 51% ($n = 86$) of the articles. The remaining 81 did not mention how they dealt with missing data (Table 3). The most common approach used was complete case analysis (62%; $n = 53$) – which was implicitly determined if articles mentioned they excluded individuals due to missing data. Among the 81 articles that did not mention any use of missing data methods, 61 explicitly indicated to have used multivariable regression models to estimate PSs using standard software (R, SAS, SPSS and STATA). Therefore, we assumed that these articles must have used complete case analysis as this is the default method for handling missing data in these analytic software. Going by this, the number of articles that used complete case analysis would be 114. Some articles also used *ad hoc* methods that are likely to result in biased estimates like: imputation to most common category ($n = 4$); mean value imputation ($n = 3$); LOCF ($n = 2$); missing indicator method ($n = 1$); and truncation which was used in a cohort study ($n = 1$) where data were utilized only up to the earliest time when the first dropout was experienced, and all the follow-up data beyond this were excluded for everybody.

19 articles (11%) used more appropriate methods like: estimation of PS by various patterns of missing data ($n = 3$); and MI, which was used in 16 of the articles among those that reported missing data methods. Among the articles that used MI; eight (50%) reported the number of imputed datasets – with the least being five and maximum 2000, and 12 (75%) explained the variables that were included in the imputation models. No article reported diagnostics on the plausibility of multiple imputed values. After conducting MI, only five articles (31%)

Table 3. Summary of missing data methods.

Missing data methods used	Number of papers				
	2010–2011	2012–2013	2014–2015	2016–2017	Total
Methods not mentioned	11 (52%)	20 (56%)	25 (45%)	25 (45%)	81 (49%)
Methods mentioned	10 (48%)	16 (44%)	30 (55%)	30 (55%)	86 (51%)
Complete case	10 (100%)	11 (31%)	16 (29%)	16 (29%)	53 (62%)
Multiple imputation	0 (0%)	2 (6%)	5 (9%)	9 (16%)	16 (19%)
Pattern mixture	0 (0%)	1 (3%)	1 (2%)	1 (2%)	3 (3%)
Imputation to most common category	0 (0%)	0 (0%)	4 (7%)	0 (0%)	4 (5%)
Simple imputation	0 (0%)	0 (0%)	2 (4%)	1 (2%)	3 (3%)
Imputation (type not specified)	0 (0%)	0 (0%)	2 (4%)	1 (2%)	3 (3%)
LOCF	0 (0%)	1 (3%)	0 (0%)	1 (2%)	2 (2%)
Truncation	0 (0%)	0 (0%)	0 (0%)	1 (2%)	1 (1%)
Missing indicator	0 (0%)	1 (3%)	0 (0%)	0 (0%)	1 (1%)

Some of the percentages did not add up to 100% due to rounding off. The percentages for missing data methods are based on the number of articles that mentioned use of methods per reporting time period.
LOCF: Last observation carried forward.

reported on how PS and outcome analysis proceeded: four articles indicated to have estimated PS and outcome analysis per imputed dataset then pooled estimates of treatment effectiveness using Rubin rules; and one article averaged PS across the imputed datasets – then used this in the adjusted analysis for all the imputed datasets with effects also pooled using Rubin rules.

Across the four reporting time periods, the proportion of articles using complete case analysis was higher in the articles published in 2010–2013 compared with the use in articles published in 2014–2017. The proportion of articles using MI increased across the four reporting time periods, from none in 2010–2011 to 16% in 2016–2017.

Missing data sensitivity analysis

The only form of sensitivity analysis that was conducted in five articles was the comparison of treatment effectiveness estimates derived from analyses of multiple imputed datasets and those from complete cases – or comparison of estimates between completers and noncompleters in cohort studies. However, no thorough sensitivity analyses aiming to rule out the possibility of MNAR were reported.

Adherence to STROBE guidelines of reporting missing data

As presented above, adherence to different aspects of STROBE guidelines pertaining to reporting of missing data were as follows: indicating reasons for missing data ($n = 12$; 7%), reporting amount of missing data ($n = 62$; 37%), and indicating missing data method(s) used ($n = 86$; 51%). Overall, only eight (5%) articles adhered to all these three aspects.

Discussion

Among 167 studies, only 86 (51%) discussed missing data issue and among them only 16 (19%) used MI methods to account correctly for the missing data. However even in these studies, reporting was incomplete as only five (31%) described how results were generated across the imputed datasets. Approximately 68% of the articles based effect estimates on complete case analysis but in the majority of the cases, this was an implicit not explicit analysis strategy. This result is consistent with the findings of Karahalios (2012) [21] and Eekhout (2012) [22] who reviewed the use of missing data methods in epidemiologic studies (though not specific to the use of PS) and also found that most researchers used complete case analysis in practice. Complete case analysis may provide unbiased estimates only if data are MCAR. However, only one of the studies provided evidence for MCAR and thus appropriately estimated PS and treatment effectiveness on complete cases. In other cases, researchers used complete case analysis but offered no explanation for why data were missing nor the underlying data generation mechanism. This use of complete case analysis may result in elimination of substantial numbers of observations and the implicit assumption of MCAR can potentially give biased results. In situations where it is not known how data were generated (as could be the case when using registry datasets), it is plausible to assume that data are MAR – then use MI in estimating

PSs. Thereafter sensitivity analysis should be used to help validate this assumption because the existence of the MNAR assumption may not be ruled out where MAR is thought to hold [26,27].

Imputation is aimed at maximizing the use of all available data. However, less principled methods like simple imputation, LOCF, imputation to most common category, truncation and missing indicator (as were used in some of the articles in this review) should be avoided [21].

This review has a number of strengths and limitations. It is the first study to systematically examine how missing data are handled with PS methods. As comparative effectiveness is often an implied concept rather than being explicitly stated, we did not restrict our search term to further include ‘comparative effectiveness’, and we were able to obtain a potentially more representative number of articles than if this was included as part of the search. However, we do acknowledge that some articles may have been excluded as searches were not conducted for the grey literature. Study selection was only carried out by one reviewer, however inclusion criteria were clearly defined and most studies were excluded as they were not clinical observational studies comparing treatment outcomes. Moreover, additional review of a random selection of included articles resulted in a very high level of agreement.

Conclusion

Researchers should pay attention to how missing data are methodologically addressed to avoid unknowingly reporting biased results. These would include: (a) identifying the most appropriate missing data mechanism (either MCAR or MAR or MNAR), and if unknown then researchers may assume data are MAR; (b) applying an appropriate missing data method based on the identified/assumed missing data mechanism and; (c) conducting sensitivity analysis in the MNAR framework if missing data mechanism was unknown. As also suggested by Karahalios (2012) [21], authors and editors should follow STROBE guidelines to increase the reliability of findings. That is, reasons for missing data should be indicated, amount of missing data and methods used to handle missing data should be reported.

Executive summary

- Majority of articles do not report how missing data are addressed in propensity score estimation.
- Among articles that report missing data methods, majority estimate propensity scores on complete cases.
- The use of complete case analysis in propensity score estimation may result in biased findings if data are not missing completely at random.
- As the underlying data generation mechanism may often be unknown at the time of analysis, researchers are encouraged to assume that data are missing at random and use imputation techniques (then estimate propensity scores on imputed datasets) followed by sensitivity analyses.

Acknowledgements

The authors would like to thank N Roberts, an information specialist with Oxford University, for her guidance in constructing the search strategies and overall use of the search engines.

Author contributions

L Malla did an initial draft of this manuscript with the support of R Perera-Salazar, E McFadden, M Ogero, K Stepniewska and M English. Thereafter, all authors edited subsequent versions and approved the final copy.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/cer-2017-0071

Financial & competing interests disclosure

The authors are grateful for the funds from the Wellcome Trust (#097170) that support M English through a fellowship and additional funds from a Wellcome Trust core grant awarded to the KEMRI-Wellcome Trust Research Programme (#092654) that supported this work. L Malla is supported by a Nuffield Department of Medicine Prize DPhil Studentship and Clarendon Scholarship (Oxford University). The funders had no role in drafting or submitting this manuscript. 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. Also the authors declare they have no competing interests.

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

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

- 1 Ye C, Beyene J, Browne G, Thabane L. Estimating treatment effects in randomized controlled trials with noncompliance: a simulation study. *BMJ Open* 4(6), e005362 (2014).
- 2 Kausto J, Solovieva S, Virta LJ, Viikari-Juntura E. Partial sick leave associated with disability pension: propensity score approach in a register-based cohort study. *BMJ Open* 2(6), e001752 (2012).
- 3 West SG, Duan N, Pequegnat W *et al.* Alternatives to the randomized controlled trial. *Am. J. Public Health* 98(8), 1359–1366 (2008).
- 4 Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. *Value Health* 12(8), 1044–1052 (2009).
- 5 Rosenbaum PR, Silber JH. Matching and thick description in an observational study of mortality after surgery. *Biostatistics* 2(2), 217–232 (2001).
- 6 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci.* 25(1), 1–21 (2010).
- 7 Thoemmes FJ, Kim ES. A systematic review of propensity score methods in the social sciences. *Multivariate Behav. Res.* 46(1), 90–118 (2011).
- **Provides a current description of how propensity score (PS) methodological aspects are used.**
- 8 Zakrisson TL, Austin PC, Mccredie VA. A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. *Eur. J. Trauma Emerg. Surg.* doi:10.1007/s00068–017–0786–6 (2017) (Epub ahead of print).
- **Provides a current description of how PS methodological aspects are used.**
- 9 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *JSTOR* 70, 41–55 (1983).
- **Describes how PS approach was developed.**
- 10 Toh S, Garcia Rodriguez LA, Hernan MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. *Pharmacoevid. Drug Saf.* 21(Suppl. 2), 13–20 (2012).
- 11 Lee KJ, Simpson JA. Introduction to multiple imputation for dealing with missing data. *Respirol.* 19(2), 162–167 (2014).
- 12 Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat. Methods Med. Res.* 16(3), 259–275 (2007).
- 13 Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat. Methods Med. Res.* 25(1), 188–204 (2016).
- **Compares various methods of estimating PS after multiple imputation.**
- 14 Rubin DB, Schenker N. Multiple imputation in healthcare databases: an overview and some applications. *Stat. Med.* 10(4), 585–598 (1991).
- 15 Mayer B, Puschner B. Propensity score adjustment of a treatment effect with missing data in psychiatric health services research. *Epidemiol. Biostatistics Public Health* 12(1), doi:http://dx.doi.org/10.2427/10214 (2015) (Epub ahead of print).
- 16 Nakai M, Ke W. Review of the methods for handling missing data in longitudinal data analysis. *Int. J. Math* 5(1), 1–13 (2011).
- 17 Yao XI, Wang X, Speicher PJ *et al.* Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *J. Natl Cancer Inst.* 109(8), doi:10.1093/jnci/djw323 (2017) (Epub ahead of print).
- 18 Kang H. The prevention and handling of the missing data. *Korean J. Anesthesiol.* 64(5), 402–406 (2013).
- 19 Fielding S, MacLennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* 9, 51 (2008).
- 20 Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Med. Res. Methodol.* 14, 118 (2014).
- **Reviews how missing data are addressed, though not with respect to the use of PSs.**
- 21 Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Med. Res. Methodol.* 12, 96 (2012).
- **Reviews how missing data are addressed, though not with respect to the use of PSs.**

- 22 Eekhout I, De Boer RM, Twisk JW, De Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology* 23(5), 729–732 (2012).
- **Reviews how missing data are addressed, though not with respect to the use of PSs.**
- 23 Vandembroucke JP, Von Elm E, Altman DG *et al.* Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology* 18(6), 805–835 (2007).
- 24 Mchugh ML. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)* 22(3), 276–282 (2012).
- 25 Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J. Clin. Epidemiol.* 62(10), 1006–1012 (2009).
- 26 Molenberghs G, Beunckens C, Sotito C. Every missing not at random model has a missingness at random counterpart with equal fit. *J. R. Stat. Soc.* 70(2), 371–388 (2008).
- 27 Molenberghs G, Thijs H, Jansen I *et al.* Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 5(3), 445–464 (2004).