OXFORD

JNCI J Natl Cancer Inst (2017) 109(8): djw323

doi: 10.1093/jnci/djw323
First published online March 22, 2017
Review

REVIEW

Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies

Xiaoxin I. Yao*, Xiaofei Wang*, Paul J. Speicher, E. Shelley Hwang, Perry Cheng, David H. Harpole, Mark F. Berry, Deborah Schrag, Herbert H. Pang

Affiliations of authors: School of Public Health, Li Ka Shing Faculty of Medicine, Hong Kong SAR, China (XIY, PC, HHP); Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, (XW, HHP); Duke University Medical Center and Duke Cancer Institute, Durham, NC, (PJS, ESH, DHH); Stanford University Medical Center, Stanford, CA, (MFB); Dana-Farber/Harvard Cancer Center, Boston, MA (DS).

*Authors contributed equally to this work.
Correspondence to: Xiaofei Wang, PhD, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Road, Durham, NC 27705 (e-mail: xiaofei.wang@duke.edu).

Abstract

Background: Propensity score (PS) analysis is increasingly being used in observational studies, especially in some cancer studies where random assignment is not feasible. This systematic review evaluates the use and reporting quality of PS analysis in oncology studies.
Methods: We searched PubMed to identify the use of PS methods in cancer studies (CS) and cancer surgical studies (CSS) in major medical, cancer, and surgical journals over time and critically evaluated 33 CS published in top medical and cancer journals in 2014 and 2015 and 306 CSS published up to November 26, 2015, without earlier date limits. The quality of reporting in PS analysis was evaluated. It was also compared over time and among journals with differing impact factors. All statistical tests were two-sided.
Results: More than 50% of the publications with PS analysis from the past decade occurred within the past two years. Of the studies critically evaluated, a considerable proportion did not clearly provide the variables used to estimate PS (CS 12.1%, CSS 8.8%), incorrectly included non baseline variables (CS 3.4%, CSS 9.3%), neglected the comparison of baseline characteristics (CS 21.9%, CSS 15.6%), or did not report the matching algorithm utilized (CS 19.0%, CSS 36.1%). In CSS, the reporting of the matching algorithm improved in 2014 and 2015 ($P = .04$), and the reporting of variables used to estimate PS was better in top surgery journals ($P = .008$). However, there were no statistically significant differences for the inclusion of non baseline variables and reporting of comparability of baseline characteristics.
Conclusions: The use of PS in cancer studies has dramatically increased recently, but there is substantial room for improvement in the quality of reporting even in top journals. Herein we have proposed reporting guidelines for PS analyses that are broadly applicable to different areas of medical research that will allow better evaluation and comparison across studies applying this approach.

Randomized controlled trials are the gold standard in clinical research but are difficult to conduct because of many practical considerations, particularly for treatments that include surgical interventions. Propensity score (PS) analysis of observational studies is an alternative method of estimating causal treatment effects for clinically important questions in observational

Received: July 8, 2016; **Revised:** October 30, 2016; **Accepted:** December 6, 2016

studies, and well-designed observational studies can also help enhance and complement the findings of randomized studies (1). Although observational studies cannot be regarded as a replacement for randomized studies, data generated from large observational cohorts have been used to evaluate important clinical questions where data from randomized trials are limited or do not exist (2,3). Observational studies also tend to have lower barriers and cost to subject recruitment, which may accelerate participation and accrual.

PS analysis is a causal inference technique for treatment effect estimation in observational studies by accounting for the conditional probability of treatment selection, thus allowing for reduction of bias when comparing interventions between treatment arms (4,5). This approach offers researchers the ability to better understand the potential effect of medical interventions and treatments. The use of PS analysis has grown dramatically in the last decade with wider availability of large databases such as Surveillance, Epidemiology, and End Results-Medicare (SEER-Medicare), Centers for Medicare and Medicaid Services (CMS), and the National Cancer Data Base (NCDB).

Despite its practicality, PS analysis also presents analytical and interpretation challenges (6–9). Importantly, the quality of reporting in PS analysis by studies can be variable, particularly because there are currently no standard reporting guidelines. Proper analyses and reporting can ensure that published results of PS analyses are reproducible, which in recent years has been recognized as a crucial element for high-quality research (10). Lack of consistency in reporting study results also has implications for those who plan to perform systematic review or meta-analyses (11).

The purpose of this study was to evaluate the reporting quality of PS analysis in cancer and cancer surgical studies by performing a systematic review of publications in top medical and surgery journals. We sought to highlight the challenges and issues associated with reporting PS analyses and to investigate evolving trends by publication year. Finally, we aimed to develop a set of reporting guidelines that could be used to help standardize future work.

Methods

Search Strategy and Study Selection

A three-part literature search of PS analysis in cancer and cancer surgical studies was conducted in the MEDLINE database using PubMed. Two primary cohorts of publications were created. The first cohort of propensity score cancer studies (CS) was created using a systematic search using the key words cancer and propensity score, propensity matched, or propensity analysis across the top 10 general medical journals and top 15 cancer journals (based on Web of Knowledge impact factors, listed in the Supplementary Methods, available online, and searched on December 28, 2015). Articles reported between 2014 and 2015 were identified. Comments, meta-analyses, reviews, and studies not focusing on cancer were excluded. The second cohort of propensity score cancer surgical studies (CSS) was created with a similar search using the same key words but with an additional key word, "surgery," among studies published through November 26, 2015, and without date and journal limits (date of search: December 11, 2015). A broader criterion for CSS cohort would allow us to compare the quality of reporting over time and among surgery journals of differing impact factors. Studies were eligible for inclusion if their primary question involved surgery among cancer patients, including both comparisons between surgical and non surgical treatment as well as comparisons between different types of surgery. Studies that involved quality of life or cost burden as the primary outcome, that did not have the full text available for review, or that were classified as comments, meta-analyses, reviews, or protocols were excluded. A third analysis was then performed examining time trends of cancer studies and cancer surgical studies utilizing PS analysis among high-impact journals between 2000 and 2015, using similar search criteria applied to the top 10 general medical journals, top 15 cancer journals, and top 15 surgery journals.

Titles and abstracts for all articles were screened in duplicate by two of the authors (XIY and PC) to independently render decisions regarding inclusion of each article. When consensus could not be reached, the two investigators reconciled the difference through reappraisal of the full text or after review by a third investigator (HP).

Data Extraction

Articles meeting eligibility criteria were included for data abstraction. Study characteristics recorded were cancer type, number of patients enrolled, number and type of treatments, study design, study end point and analysis, publication date, and journal. PS elements extracted included PS methods used in the estimation of treatment effect, variables used in PS estimation, whether any non baseline variables were included, the comparability of baseline characteristics in PS analyses, and the total number of subjects included in the matched analysis and matching algorithm utilized (if PS matching study). The evaluation of the matching algorithm was abstracted, including distance metric (greedy nearest-neighbor matching, greedy matching within specified caliper distances, greedy matching by digit, greedy matching without distance metric specified, and optimal matching), matching ratio, the use of replacement, and the method used to assess comparability of baseline characteristics between matched groups (12,13). The reporting of the assumptions of no unmeasured confounders and sufficient overlap in the propensity scores distribution, as well as goodness-of-fit of the model, were also abstracted (4,14,15). The reporting of variables used in PS estimation was classified as "yes," "no," or "unclear." Studies were classified as "yes" if the variables were listed out or clearly defined and were classified as "no" if the variables were neither mentioned in the text, tables, nor appendices/Supplementary Materials (available online). Otherwise, the study was classified as "unclear" if the variables were not clearly reported (eg, the variables were reported as "all relevant covariates" or "all covariates potentially predictive of treatment" without any clear definition or statement). If the answer of reporting of variables used in PS estimation was "no" or "unclear," the item whether non baseline variables were included was defined as "not evaluable". The reporting of the matching algorithm was recorded as "yes" when the method used to form matched sets of subjects was stated (eg, greedy matching, optimal matching). Other aspects such as completeness of follow-up and accuracy of end point assessment were not taken into consideration.

Variables related to reporting of PS estimation, comparability of baseline characteristics, and matching algorithm were collected and verified by two authors (XIY and PC) and then confirmed by a third author (HP). Inter-rater agreement was assessed for four variables (PS methods used in the estimation of treatment effect, cancer type, variables used to estimate the PS, and matching algorithm) with 60 randomly selected articles by two data extractors, XIY

and PC, separately. The corresponding Cohen's κ coefficients for these four variables were 0.95, 1.00, 0.82, and 1.00, respectively, which indicated substantial to perfect interrater agreement (16). No important discrepancies were observed.

Statistical Methods

Categorical variables for characteristics and reporting of CS and CSS were described with frequencies and percentages. Median and interquartile range were used for number of patients enrolled. To investigate evolving trends by publication year and reporting quality in differently ranked journals, Fisher's exact tests were used to compare the reporting of CSS published on or before 2013 vs 2014/2015, and to compare the reporting quality of top 15 vs non-top 15 surgery journals from CSS. Ninety-five percent confidence intervals for difference in proportions were also reported. All reported *P* values were two-sided, and a *P* value of less than .05 was considered statistically significant. Statistical analysis was performed using R version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Study Selection and Time Trends

We identified 37 cancer-focused studies involving PS methods reported in top medical/cancer journals between 2014 and 2015, of which 33 met the inclusion criteria (18 in *Journal of Clinical Oncology*, six in *Journal of the National Cancer Institute*, four in *BMJ*, three in *The Lancet Oncology*, and two in *JAMA*) (Figure 1A). For cancer surgical studies, 505 citations were identified via PubMed (Figure 1B; Supplementary Table 1, available online). Four hundred eighty articles were selected after screening titles and abstracts, and 306 eligible articles were included on the basis of their full text. The most common reason for exclusion among these publications was that the primary question did not involve surgery. Our study closely followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist as illustrated in Supplementary Table 2 (available online).

Time trends of cancer studies reporting use of PS analysis between 2000 and 2015 among the top medical/cancer and surgery journals are presented in Figure 2. The number of cancer studies using PS has grown markedly in recent years. The last two years alone accounted for more than 50% of the total papers published in the past decade. The top three journals were *Journal of Clinical Oncology* (n = 37), *Journal of Thoracic and Cardiovascular Surgery* (n = 29), and *Annals of Surgical Oncology* (n = 28). A fairly similar time trend was also found in the CSS cohort (Supplementary Figure 1, available online), with 172 articles published in 2014 and 2015 and 134 articles published up through 2013.

Characteristics of Included Studies

Table 1 describes the main characteristics of studies reviewed. The top three cancer types in CS and CSS were gastrointestinal cancer, followed by lung cancer, then genitourinary cancer. CS tended to enroll more patients than CSS, with a median of 4515 (Q1 to Q3, 1392 to 20600) vs 699 (Q1 to Q3, 307 to 2783). Among the PS matching papers, 19.0% of CS did not mention the proportion of matched sample size, that is, sample size after matching over sample size before matching, while it was only 5.5% for CSS. In addition, the overall matched proportion was less than 50% in eight CS (38.1%) and 115 CSS (52.5%). For the matched proportion of the targeted treatment group, see Supplementary Table 3 (available online). Most of the articles reviewed compared two treatments (97.0% and 90.5%, respectively). There was only one CS (3%) involving more than two treatment groups, compared with 29 CSS (9.5%). The frequencies of different PS methods (ie, propensity score matching [PSM], propensity score weighting [PSW], propensity score stratification (PSS), covariate adjustment using propensity score [CAPS], and more than one type of PS methods) as reported across the studies are available in Table 1. Overall, 16 CS (48.5%) and 207 CSS (67.6%) utilized PSM; six CS (18.2%) and 21 CSS (6.9%) utilized more than one type of PS methods, for example, both PSM and PSW. A summary of the study design and the study end point and analysis can be found in Supplementary Tables 4 and 5 (available online). The majority of the articles focused on survival or time-to-event outcomes (81.8% and 73.2%, respectively), followed by dichotomous or discrete outcomes (18.2% and 31.4%, respectively) and continuous outcomes (3.0% and 12.7%, respectively). More than one primary outcome was of interest in some studies; therefore the total percentages are over 100.

Reporting of PS Methodology

There were four CS (12.1%) and 27 CSS (8.8%) that did not provide the variables used in PS estimation clearly (Table 2). In addition, one CS (3.4%) and 26 (9.3%) CSS incorrectly included non baseline variables. In 32 CS and 275 CSS, which involved PSM, PSW, or PSS methods, seven (21.9%) and 43 (15.6%) of the studies, respectively, did not report the comparability of baseline characteristics between treated and untreated subjects in the matched, weighted, or stratified sample. For those 25 CS and 232 CSS that did report baseline comparisons, two CS (8.0%) and 31 CSS (13.4%) found imbalanced characteristics. Among the 21 CS and 219 CSS that utilized PSM, four CS (19.0%) and 79 CSS (36.1%) did not report the matching algorithm. Four CS (19.0%) and 97 CSS (44.3%) did not report the distance metric. The matching ratio was reported by all CS, but not reported in eight CSS (3.7%). Fifteen CS (71.4%) and 188 CSS (85.8%) did not report whether replacement was used. Sixteen CS (76.2%) and 195 CSS (89.0%) assessed the comparability of baseline characteristics between matched groups, while four CS (19.0%) and 93 CSS (42.5%) did not clearly state the method used. Thirteen CS (39.4%) and 73 CSS (23.9%) discussed or mentioned the assumption of no unmeasured confounders. Four CS (12.1%) and 14 CSS (4.6%) described the distribution of propensity scores among the compared treatment groups. One CS (3.0%) and 28 CSS (9.2%) assessed the goodness-of-fit of the PS estimation model, while two CS (6.1%) and 14 CSS (4.6%) assessed the goodness-of-fit of the outcome model.

The reporting of matching algorithm in CSS improved in the last two years ($P = .04$) (Table 3). However, there were no statistically significant improvements for the reporting of variables used ($P = .23$), inclusion of non baseline variables ($P = .22$), and reporting of comparability of baseline characteristics ($P = .14$). The reporting of variables used to estimate PS was better in the top 15 surgery journals ($P = .008$) (Table 3). However, there were no statistically significant differences for the inclusion of non-baseline variables ($P = .49$), reporting of comparability of baseline characteristics ($P = .43$), and reporting of matching algorithm ($P = .31$).

Discussion

The number of manuscripts utilizing PS methods in cancer and cancer surgical journals has rapidly increased in recent years,

REVIEW

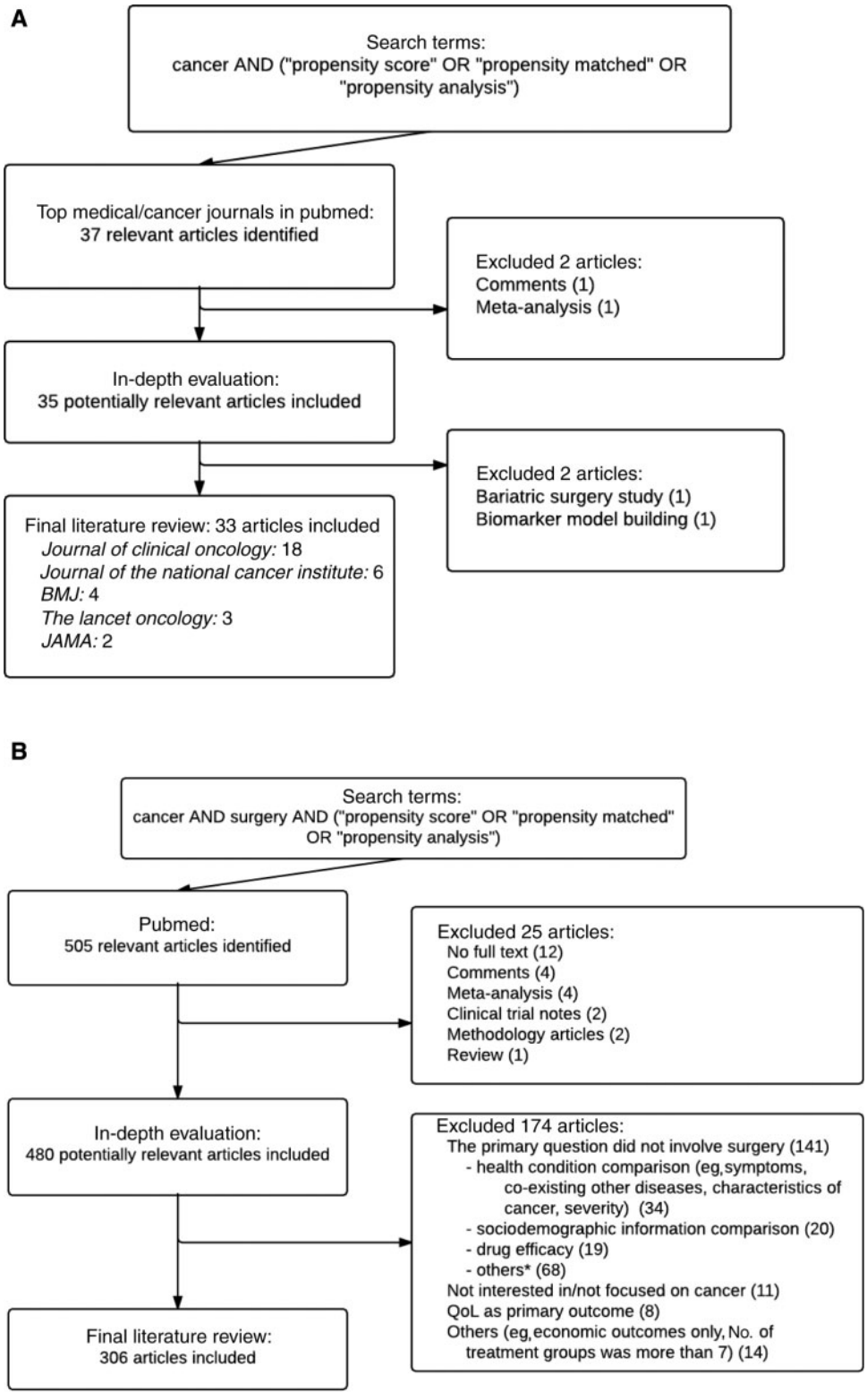

Figure 1. Process of literature search: (A) cancer studies in top medical and cancer journals and (B) cancer surgical studies. *Other reasons are given in Supplementary Table 1 (available online).

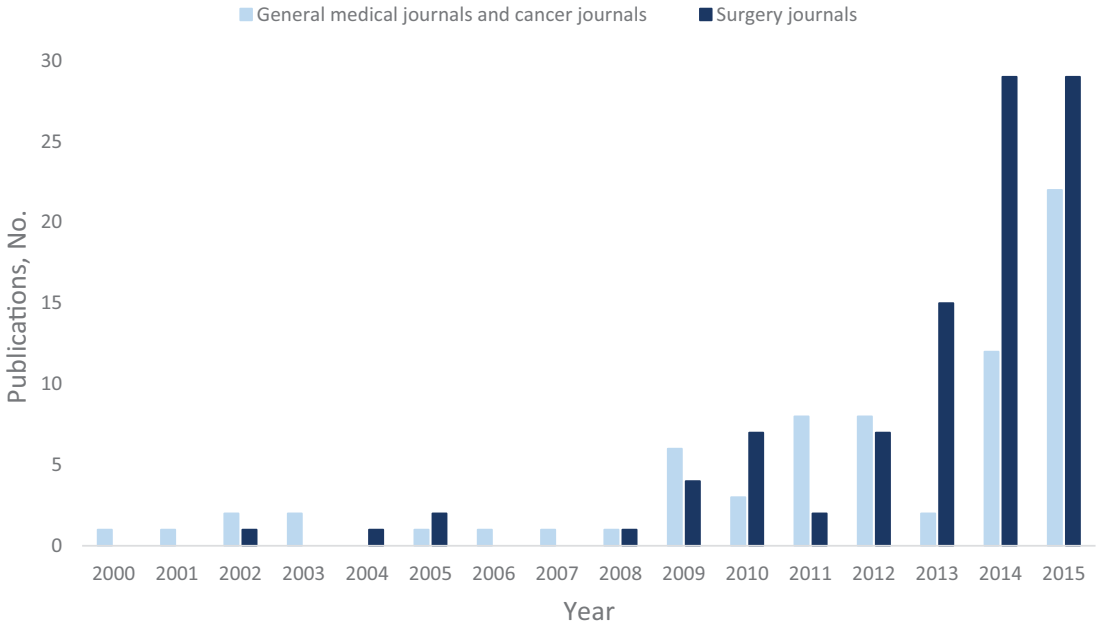

Figure 2. Publication trends in cancer studies reporting use of propensity score analysis in high-impact medical/cancer and surgery journals.

likely driven in large part by the increasing availability of large databases, such as SEER-Medicare, CMS, and NCDB. In this evaluation of the quality of PS reporting in over 300 cancer-related observational studies, we demonstrate that essential methodological information is often not reported. Our results indicate that the quality of PS reporting in cancer studies requires substantial improvement, even in high–impact factor journals. Our findings clearly support the need for reporting guidelines for PS analyses.

Inadequate reporting of PS analyses could have severe consequences on the interpretation of a study's findings and potentially impact either subsequent research or even clinical care. First, the design of future clinical studies and patient management can be informed by the results of these large PS analyses. Second, inadequate reporting of variables included in the PS model, not reporting the matching algorithm utilized, and inconsistent reporting of variables in the "Methods" and "Results" sections represent poor data provenance and can result in problems with study reproducibility and interpretation. For example, a considerable number of studies included in our analysis, 21.9% CS and 15.6% CSS, did not report the comparability of measured baseline characteristics after the application of the PS methods, which makes it difficult to judge the appropriateness and effectiveness of the PS analysis and its results (17). Moreover, comparisons of balance among measured baseline variables should be performed with proper methods for matched samples (8,9). The recommended methods to assess the comparability of baseline characteristics between matched groups include the standardized difference (<10%) and the C-statistic (close to 0.5) (13,18).

Another area of concern that we identified relates to limited reporting of matched sample size proportions. Incomplete matching can bias the research findings, especially when a sizable proportion of subjects in the treatment group are excluded after matching. The inference of treatment effects derived from a limited subset of subjects can systematically differ from the target population for inference, which means that the estimated average treatment effect on the treated can be biased. When choosing the best matching algorithm (as well as the distance metric, ratio, and the use of replacement), there usually exists a trade-off between the bias from sizable sample loss and residual confounding from the inclusion of poorly matched subjects (19). Therefore, the reporting of matched sample size proportions and whether the covariate distribution after matching is subsequently retained for the treatment comparison have important implications regarding the interpretation of results.

Despite the broad usage, PS analysis is limited by its inability to control for unmeasured confounders and variables measured with error (14,15). However, as we have noted in the "Results" section, these assumptions were often not reported. Also, a low proportion of studies reported the amount of overlap in the propensity score distributions among the matched treatment groups. Sufficient overlap in the distribution of propensity scores among the comparison groups is also an important assumption in PS analysis to ensure valid causal inference (20–22).

Finally, lack of consistency in reporting study methods and results gives rise to difficulty in performing systematic reviews or meta-analyses. Prior studies have shown that inappropriate and poor reporting can lead to misleading results and can waste valuable resources (10,23–25). The lack of consistency across the reporting of PS analyses calls for more standardized and reproducible reporting of study methods and results, especially considering the dramatic increase in the recent use of these statistical methods in studies of cancer patients.

To improve consistency and reproducibility, we propose a set of guidelines and recommendations (a concise list shown in Table 4, with an expanded version in Supplementary Table 6, available online) to ensure comprehensive, complete, and clear reporting in PS analyses. Based on the systematic review, items that could impact the reproducibility and interpretability of a PS analysis were generated. These items were then integrated with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) categories for reporting observational

Table 1. Characteristics of included studies

Variables	Cancer studies* No. of studies/total No. (%)	Cancer surgical studies No. of studies/total No. (%)
Cancer type		
Gastrointestinal cancer	7/33 (21.2)	118/306 (38.6)
Lung cancer	7/33 (21.2)	89/306 (29.1)
Genitourinary cancer	6/33 (18.2)	67/306 (21.9)
Breast cancer	5/33 (15.2)	20/306 (6.5)
Thyroid cancer	0	5/306 (1.6)
Head and neck cancers	0	3/306 (1.0)
Hematopoietic and lymphoid cancers	3/33 (9.1)	0
Skin cancer	1/33 (3.0)	0
Nervous system cancer	0	1/306 (0.3)
Advanced/metastatic cancer	3/33 (9.1)	0
Others	1/33 (3.0)	3/306 (1.0)
No. of treatment groups		
2	32/33 (97.0)	277/306 (90.5)
3	1/33 (3.0)	21/306 (6.9)
≥4	0	8/306 (2.6)
No. of patients enrolled		
Median	4515	699
Q1 to Q3	1392 to 20600	307 to 2783
Not mentioned	0	3/306 (1.0)
Propensity score methods type		
Propensity score matching	16/33 (48.5)	207/306 (67.6)
Propensity score weighting	9/33 (27.3)	20/306 (6.5)
Propensity score stratification	1/33 (3.0)	27/306 (8.8)
Covariate adjustment using propensity score	1/33 (3.0)	31/306 (10.1)
More than one type	6/33 (18.2)	21/306 (6.9)
The proportion of matched sample size†		
<25%	2/21 (9.5)	35/219 (16.0)
25%–<50%	6/21 (28.6)	80/219 (36.5)
50%–<75%	3/21 (14.3)	53/219 (24.2)
75%–<100%	6/21 (28.6)	14/219 (6.4)
Not mentioned	4/21 (19.0)	12/219 (5.5)
Only reported the matched sample	0	25/219 (11.4)

*Cancer studies in top medical and cancer journals in 2014 and 2015. Nine studies were included in both cancer studies and cancer surgical studies.
†The reporting of the proportion of matched sample size was evaluated in studies utilizing matching.

studies (26). We believe that following the proposed guidelines should substantially improve the reporting quality of PS analysis. While cancer and cancer surgery studies were used to generate these guidelines because of the authors' combined expertise and the growing use of PS analyses in the oncology literature in recent years (Supplementary Figure 2, available online), the recommendations put forward apply to a broad range of observational research. Of these recommendations, perhaps the most important include the need to state the specific PS method(s) used, to specify the model used in PS estimation including the variables used, to describe comparisons of baseline characteristics, and to explicitly specify the method used to form matched sets of subjects, if matching is used. The set of guidelines should not be viewed as a comprehensive manual for conducting proper statistical analyses using PS methods, but it should be viewed as a tool for consistent reporting, reproducibility, and interpretation of PS analysis. These guidelines are a first step toward improving PS analysis reporting and can be further refined.

While consistent reporting, reproducibility, and interpretation of PS analysis are our primary focuses, we would like to highlight a few references related to the proper use of PS analysis. A few studies reviewed made comparisons among more than two treatment groups. In this instance, special approaches are needed to fit propensity scores. Investigators should, for example, use a multinomial logistic regression, a multinomial probit model, or a series of binary probits to estimate propensity scores (19,27). PS analysis has been occasionally used in case-control studies. In this review, we have one example of this study design. When applying PS analysis to case-control study, the investigator should consider the impact of artifactual effect modification and residual confounding on the study findings. More complete discussion on this topic can be found in Månsson et al. (2007) (28). In this paper, the studies investigated did not involve the effect of time-varying treatment. If the effect of time-varying treatment is of interest, special considerations of time-varying covariates for PS estimation will be necessary (29,30). When using PSS method, the number of strata should be chosen based on an analysis of the rate at which the number of strata should increase with sample size to reduce residual confounding (6). In general, it has been found that PSW is less prone to model misspecification than CAPS (31). In addition, when the number of outcomes is low relative to the number of confounders, confounder control through use of PS analysis provides less biased and more precise estimated treatment effects than multivariable logistic regression (32). PS analysis allows for the estimation of marginal treatment effect, an average effect at the

Table 2. Reporting of propensity score analysis in included studies

Variables	Cancer studies* No. of studies/total No. (%)	Cancer surgical studies No. of studies/total No. (%)
Variables used to estimate the PS		
Yes	29/33 (87.9)	279/306 (91.2)
No/unclear	4/33 (12.1)	27/306 (8.8)
Inclusion of non baseline variables†		
Yes	1/29 (3.4)	26/279 (9.3)
No	28/29 (96.6)	253/279 (90.7)
Comparability of baseline characteristics‡		
Yes	25/32 (78.1)	232/275 (84.4)
No	7/32 (21.9)	43/275 (15.6)
Matching algorithm§		
Yes	17/21 (81.0)	140/219 (63.9)
No	4/21 (19.0)	79/219 (36.1)
Distance metric§		
Greedy nearest neighbor matching	4/21 (19.0)	46/219 (21.0)
Greedy matching within specified caliper distances	8/21 (38.1)	49/219 (22.4)
Greedy matching by digit	4/21 (19.0)	19/219 (8.7)
Greedy matching without distance metric specified	0	18/219 (8.2)
Optimal matching	1/21 (4.8)	8/219 (3.7)
Not reported	4/21 (19.0)	79/219 (36.1)
Matching ratio§		
Yes	21/21 (100)	211/219 (96.3)
No	0	8/219 (3.7)
Use of replacement§		
With replacement	1/21 (4.8)	4/219 (1.8)
Without replacement	5/21 (23.8)	27/219 (12.3)
Not reported	15/21 (71.4)	188/219 (85.8)
Method to assess comparability of baseline characteristics between matched groups§		
Standardized difference	6/21 (28.6)	31/219 (14.2)
C-statistic	0	17/219 (7.8)
Absolute difference	1/21 (4.8)	1/219 (0.5)
Paired test‖	2/21 (9.5)	20/219 (9.1)
Independent sample test¶	3/21 (14.3)	31/219 (14.2)
Regression	0	2/219 (1.0)
Assessed but method not reported	4/21 (19.0)	93/219 (42.5)
Not assessed	5/21 (23.8)	24/219 (11.0)
Imbalanced baseline characteristics#		
Yes	2/25 (8.0)	31/232 (13.4)
No	23/25 (92.0)	201/232 (86.6)

*Cancer studies in top medical and cancer journals in 2014 and 2015. Nine studies were included in both cancer studies and cancer surgical studies. PS = propensity score.

† The reporting of whether non baseline variables were included was not evaluable if the answer of reporting of variables used to estimate the PS was "no/unclear".

‡ The reporting of comparability of baseline characteristics in PS analyses was evaluated in studies utilizing matching, weighting, or stratification.

§ The reporting was evaluated in studies utilizing matching.

‖ The statistical test for paired or matched sample, eg, paired *t* tests, Wilcoxon signed rank test, and McNemar's test.

¶ The statistical test for independent sample, eg, unpaired *t* tests, Wilcoxon rank sum test, chi-square test, and Fisher's exact test.

The reporting of whether baseline characteristics were imbalanced was evaluated in studies reporting comparability of baseline characteristics.

population level, whereas multivariable regression yields conditional treatment effect, an average effect at the individual level (33). When outcomes are binary or time-to-event in nature, which are of primary interest in most studies in this review, the marginal odds ratio or hazard ratio would generally be closer to the null than the conditional effect, while the statistical significance of the estimates would usually be similar (34–36). Moreover, Knol et al. (2012) offered some suggestions in reporting the effect interaction and modification (37).

In this study, we analyze the existing oncology literature to develop a set of novel guidelines for reporting PS analyses. In an era of rapidly increasing use of PS techniques, such guidelines are essential to promote study reproducibility and to responsibly inform patient care and future prospective research. This is a large-scale study to scrutinize and evaluate the reporting of PS analyses. Importantly, the descriptions and analyses of the quality of PS reporting are not unique to cancer studies, but our guidelines serve as a framework for evaluating the quality of reporting in PS analyses in other areas as well. Our reporting guidelines can also serve as an evaluation tool in the peer review process for assessing future research involving PS analysis.

Despite notable strengths, our study has some limitations. First, the literature search utilized was based only on literature contained within the MEDLINE database and reported in

Table 3. Reporting quality of cancer surgical studies by publication year (on or before 2013 vs 2014/2015) and by journal ranking (non-top 15 vs top 15)

Variables	Cancer surgical studies			Cancer surgical studies		
	On or before 2013	2014/2015	P*	Non-top 15 journals	Top 15 journals	P*
Variables used to estimate the PS						
Yes, No. (%)	119 (88.8)	160 (93.0)	.23	208 (88.9)	71 (98.6)	.008
No/unclear, No. (%)	15 (11.2)	12 (7.0)		26 (11.1)	1 (1.4)	
Difference in proportions (95% CI)	–	4.2 (−2.7 to 11.8)		–	9.7 (1.7 to 14.8)	
Inclusion of non baseline variables						
Yes, No. (%)	8 (6.7)	18 (11.3)	.22	18 (8.7)	8 (11.3)	.49
No, No. (%)	111 (93.3)	142 (88.8)		190 (91.3)	63 (88.7)	
Difference in proportions (95% CI)	–	4.5 (−3.3 to 11.7)		–	2.6 (−5.1 to 13.4)	
Comparability of baseline characteristics						
Yes, No. (%)	95 (80.5)	137 (87.3)	.14	179 (83.3)	53 (88.3)	.43
No, No. (%)	23 (19.5)	20 (12.7)		36 (16.7)	7 (11.7)	
Difference in proportions (95% CI)	–	6.8 (−2.4 to 16.4)		–	5.1 (−7.3 to 13.8)	
Matching algorithm						
Yes, No. (%)	47 (55.3)	93 (69.4)	.04	113 (65.7)	27 (57.4)	.31
No, No. (%)	38 (44.7)	41 (30.6)		59 (34.3)	20 (42.6)	
Difference in proportions (95% CI)	–	14.1 (0.4 to 27.6)		–	−8.3 (−25.0 to 7.7)	

*Two-sided Fisher's exact test. CI = confidence interval; PS = propensity score.

Table 4. Brief guidelines for reporting propensity score analysis

Section/topic	Item	No.*	Recommendation
Title and abstract	☐	1	Indicate the use of propensity analysis with a commonly used term in the title or the abstract
Methods			
Bias	☐	9	Describe how propensity score analysis was used to address bias
Statistical analyses	☐	12	Describe all the analytic methods, including the propensity score methods, eg, PSM, PSW, PSS, CAPS
	☐	13	Indicate the model used to estimate propensity score
	☐	14	State the variables included in the propensity score model
	☐	15	Explain the variable selection procedure for propensity score model
	☐	16	PSM: Explicitly state the matching algorithm and distance metric, indicate matching ratio (1:m matching), indicate whether sampling with or without replacement was used, describe the statistical methods for the analysis of matched data, report the package used to create matched sample, and describe methods for assessing the comparability of baseline characteristics in the matched groups
	☐	17	PSW: Describe methods for assessing the comparability of baseline characteristics in the weighted groups
	☐	18	PSS: Give the number of strata and describe methods for assessing the comparability of baseline characteristics in each stratum
	☐	19	Explain how assumption of propensity score analysis was examined
	☐	20	Explain how missing data in propensity score estimation were addressed
Results			
Participants	☐	25.4	PSM: Report the sample size for each treatment group before and after matching
Patient characteristics	☐	28	Describe the distribution of baseline characteristics for each group before propensity score analysis
	☐	29	PSM, PSW, PSS: Describe the distribution of baseline characteristics in the matched/weighted groups or in each stratum, and describe the results of the comparability of baseline characteristics
	☐	30	Indicate number of patients with missing data for each variable of interest, especially the variables used in propensity score model
Main results	☐	32	Give propensity score analysis estimates and their precision, eg, 95% confidence interval
	☐	33	If applicable, give unadjusted estimates and/or adjusted estimates and their precision, eg, 95% confidence interval, and make clear which additional factors were adjusted for
Discussion			
Interpretation	☐	38	Discuss whether imbalance of baseline characteristics still exists, and give a cautious interpretation
Generalizability	☐	40	PSM: Discuss the possibility and potential influence of incomplete matching, especially the studies in which the matched sample size is less than 50%

*For full guidelines, refer to Supplementary Table 6 (available online). CAPs = covariate adjustment using propensity score; PSM = propensity score matching; PSS = propensity score weighting; PSW = propensity score weighting.

PubMed. However, recent studies have shown that using data sources beyond PubMed has only modest impact on the results of systematic reviews (38,39). Second, the cancer studies in top journals included in this review were limited to those published between 2014 and 2015. Despite this, the studies included accounted for more than half of the literature utilizing PS methods published to date and represent the most contemporary use of PS methods in the current literature.

In conclusion, propensity score analysis is a statistical technique commonly used to estimate causal treatment effects for clinical interventions in observational studies. The use of this analytical approach has particularly increased in clinical research involving surgical interventions. We find that current reporting is often inadequate and ambiguous, even in high-impact medical journals. Accordingly, we propose rational reporting guidelines to foster transparency and consistency and to facilitate interpretation by readers. The purpose of these guidelines is to set forth a comprehensive and clear checklist to maximize the value of research that leverages PS techniques.

Funding

This study was supported by grant R21-AG042894 from the NIH National Institute on Aging, grant P01-CA142538 from the NIH National Cancer Institute, and Health and Medical Research Fund of Hong Kong.

Notes

The funding source had no role in design of the study; the collection, analysis, or interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

We thank our colleagues, two anonymous reviewers, and the associate editor for their valuable comments and suggestions.

The authors have no conflicts of interest to declare.

REVIEW

References

1. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med*. 2000;342(25):1907–1909.
2. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312(7040):1215–1218.
3. Silverman SL. From randomized controlled trials to observational studies. *Am J Med*. 2009;122(2):114–120.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
5. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17(19):2265–2281.
6. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
7. D'Agostino RB Jr. Propensity scores in cardiovascular research. *Circulation*. 2007;115(17):2340–2343.
8. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037–2049.
9. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. 2010;25(1):1–21.
10. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166–175.
11. Phelps RM, Dearing MP, Mulshine JL. Need for uniformity in collection and reporting of data in cancer clinical trials. *J Natl Cancer Inst*. 1990;82(17):1377–1378.
12. Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *Am J Epidemiol*. 2014;179(2):226–235.
13. Franklin JM, Rassen JA, Ackermann D, et al. Metrics for covariate balance in cohort studies of causal effects. *Stat Med*. 2014;33(10):1685–1699.
14. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 1993;49(4):1231–1236.
15. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 pt 2):757–763.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
17. McMurry TL, Hu Y, Blackstone EH, et al. Propensity scores: Methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. *J Thorac Cardiovasc Surg*. 2015;150(1):14–19.
18. Normand ST, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J Clin Epidemiol*. 2001;54(4):387–398.
19. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J Econ Surv*. 2008;22(1):31–72.
20. Rubin DB. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20–36.
21. Rubin DB. On the limitations of comparative effectiveness research. *Stat Med*. 2010;29(19):1991–1995.
22. Walker A, Patrick Lauer, et al. A tool for assessing the feasibility of comparative effectiveness research. *Com Eff Res*. 2013;3:11–20.
23. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99(2):147–157.
24. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine–reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189–2194.
25. Raghav KP, Mahajan S, Yao JC, et al. From protocols to publications: A study in selective reporting of outcomes in randomized trials in oncology. *J Clin Oncol*. 2015;33(31):3583–3590.
26. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344–349.
27. McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388–3414.
28. Månsson R, Joffe MM, Sun W, et al. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol*. 2007;166(3):332–339.
29. Lu B. Propensity score matching with time-dependent covariates. *Biometrics*. 2005;61(3):721–728.
30. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.
31. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott, Williams and Wilkins; 2008.
32. Cepeda MS. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158(3):280–287.
33. Rosenbaum PR. Propensity Score. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. 2nd ed. Boston, MA: Wiley; 2005:4267–4272.
34. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32(16):2837–2849.
35. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005;58(6):550–559.
36. Sturmer T, Joshi M, Glynn RJ, et al. 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*. 2006;59(5):437–447.
37. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41(2):514–520.
38. Bramer WM, Giustini D, Kramer BM, et al. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Syst Rev*. 2013;2:115.
39. Halladay CW, Trikalinos TA, Schmid IT, et al. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. *J Clin Epidemiol*. 2015;68(9):1076–1084.