

A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines

T. L. Zakrison³ · P. C. Austin¹ · V. A. McCredie²

Received: 8 December 2016 / Accepted: 10 March 2017 / Published online: 24 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract

Background Propensity score methods are techniques commonly employed in observational research to account for confounding when estimating the effects of treatments and exposures. These methods have been increasingly employed in the acute care surgery literature in an attempt to infer causality; however, the adequacy of reporting and the appropriateness of statistical analyses when using propensity score matching remain unclear.

Objectives The goal of this systematic review is to assess the adequacy of reporting of propensity score methods, with an emphasis on propensity score matching (to assess balance and the use of appropriate statistical tests), in acute care surgery (ACS) studies and to provide suggestions for improvement for junior investigators.

Methods We searched three databases, and other relevant literature (from January 2005 to June 2015) to identify observational studies within the ACS

literature using propensity score methods (PROSPERO No: CRD42016036432). Two reviewers extracted data and assessed the quality of the studies retrieved by reviewing the adequacy of both overall reporting and of the propensity score matching methods used.

Results A total of 49/71 (69%) of studies adequately reported propensity score methods overall. Matching was the most common propensity score method used in 46/71 (65%) studies, with 36/46 (78%) studies reporting matching methods adequately. Only 19/46 (41%) of matching studies reported the balance of baseline characteristics between treated and untreated subjects while 6/46 (13%) used correct statistical methods to assess balance. There were 35/46 (76%) of matching studies that explicitly used statistical methods appropriate for the analysis of matched data when estimating the treatment effect and its statistical significance.

Conclusion We have proposed reporting guidelines for the use of propensity score methods in the acute care surgery literature. This is to help investigators improve the adequacy of reporting and statistical analyses when using observational data to estimate effects of treatments and exposures.

Electronic supplementary material The online version of this article (doi:10.1007/s00068-017-0786-6) contains supplementary material, which is available to authorized users.

✉ T. L. Zakrison
tzakrison@med.miami.edu

P. C. Austin
peter.austin@ices.on.ca

V. A. McCredie
Victoria.McCredie@sunnybrook.ca

¹ Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

² Department of Critical Care Medicine, Sunnybrook Health Sciences Center, Toronto, ON, Canada

³ Department of Surgery, University of Miami Miller School of Medicine, 1800 NW 10th Ave., Miami, FL 33136, USA

Keywords Propensity score · Matching · Trauma · Acute care · Surgery

Introduction

Enrollment of trauma and acute care surgery patients into randomized controlled trials (RCTs) can be difficult, as it is not always feasible or ethical to perform these types of studies. Thus, carefully conducted observational studies need to be used to estimate effectiveness of interventions,

if and when possible. Observational studies allow for inference about causation, in addition to association, but only after accounting for all confounding variables. However, in such studies, treatment allocation is commonly dependent on baseline characteristics, such as severity of illness, making it difficult to account for and balance all variables. In order to circumvent these limitations, surgical researchers have begun to use propensity score methods to achieve balance in measured baseline characteristics between patients receiving treatment and those who are not, in an attempt to infer causality [1]. As such, the number of studies that use propensity-based methodology has been increasing rapidly in the acute care surgery (ACS) literature. Despite the increasing use of propensity score methods in the medical literature, reporting guidelines using propensity score methods in the field of ACS do not exist. Previous systematic reviews of reporting of propensity score methodology in the cardiovascular and internal medicine literature, for example, have shown wide variability in the accurate application of propensity score methods. Recent systematic reviews of studies in the medical literature that used this methodology have even demonstrated that these were often poorly conducted and inadequately reported [5–7]. To date, no one has systematically reviewed the use of propensity score methodology in the ACS literature. Furthermore, no reporting guidelines exist to facilitate correct reporting of this methodology and aid in the interpretation of such studies in ACS journals. These would be important as ACS patients are as heterogeneous (representing trauma, emergency general surgery or surgical critical care) as the clinicians investigating outcomes (trauma surgeon vs. critical care intensivist). No other branch of surgery has such heterogeneity; thus ACS-specific guidelines would be helpful.

Four different propensity score methods are used: (1) propensity score matching (2) stratification on the propensity score, (3) covariate adjustment using the propensity score and (4) inverse probability of treatment weighting using the propensity score. All of these methods aim to compare outcomes between comparable treated and untreated subjects, so the only “effect” on outcome is the intervention received. Propensity score methods thus aim to achieve balance between exposed and control groups, with the inclusion and measurement of all covariates that may induce confounding. While propensity score matching is one of the most common methods used [2] common misconceptions and pitfalls exist [3, 4]. Recent systematic reviews of studies in the medical literature that used this methodology have even demonstrated that these were often poorly conducted and inadequately reported [5–7]. Firstly, reporting practices for propensity score methods, including matching, must highlight important details that allow other researchers to confidently judge the appropriateness of reported analyses and potentially to replicate published

findings. This includes the inclusion and justification of the use of covariates, the largest threat to bias. Balance between matched covariates must be measured appropriately to verify that confounding has indeed been reduced. Matched statistical analysis must also be applied after propensity score matching in order to avoid an overly conservative Type I error using unmatched tests. While this list is not exhaustive, these represent the fundamental principles in the appropriate use of propensity score analysis.

In the context of such findings and because of the growing interest in using propensity-score analysis in the ACS literature, the objective of the current article is to assess the adequacy of reporting of propensity score methods, with an emphasis on propensity score matching, by performing a systematic review of the ACS literature of the last 10 years.

Methods

We conducted this systematic review using a predefined protocol according to current standards [8] and adhering to PRISMA criteria [9, 10]. The Research Ethics Board of University of Miami reviewed the study and deemed it exempt from review. Our protocol was registered with PROSPERO (No: CRD42014010405; available at http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016036432).

Search strategy

We systematically searched Medline, Embase, and Web of Science from January 2005 up to July 2015, to identify ACS studies using propensity score methods (details in Supplemental Appendix). There were no restrictions on language. Comprehensive search strategies were developed for each database based on commonly used terms, keywords and controlled vocabulary terminologies. Bibliographies of relevant narrative and systematic reviews, clinical practice guidelines, as well as retrieved studies were searched for additional studies. Retrieved studies were assessed by one of the authors (T.L.Z.), who screened the titles and abstracts to identify relevant articles. Studies were included only if they were identified as observational trials using propensity score methods. Letters and articles for which only the abstract was available or reports describing only the design of the trial were excluded.

Study selection

Two reviewers (T.L.Z., V.A.M.) independently screened studies for inclusion, retrieved potentially relevant studies, and decided on study eligibility. We selected observational cohort studies of adult ACS patients published over the

past 10 years (Jan 2005–July 2015). We included all peer-reviewed observational studies ACS using propensity score methods in the literature. ACS was defined as torso trauma (any trauma from the neck to the pelvis), emergency general surgery or surgical critical care which included trauma patients within the cohort (“critical care trauma”). We excluded studies that did not include analysis of data, randomized clinical trials, case–control studies, and studies primarily analyzing cost-effectiveness or practice patterns. We also excluded studies in which the term “propensity” was used in a literary, non-statistical way. We excluded single-system orthopedic and neurosurgical trauma, pre-hospital, disaster and military trauma.

Data abstraction, evaluation and outcome measures

Two reviewers (T.L.Z.,V.A.M.) independently abstracted data and methodological characteristics from the included studies using a standard form (Supplementary Appendix 2). Disagreements were resolved by consensus and if necessary, in consultation with a third reviewer (P.C.A.). These two independent reviewers tested a data extraction form with a distinct set of five studies during a training session. There are no formal reporting guidelines in the ACS literature for the use of propensity score methods. We critically reviewed the study methodology using suggestions from previously published review articles [1, 5, 7, 11]. We judged the adequacy of reporting of propensity score methodology with the following: (a) report characteristics (trauma, emergency general surgery or critical care trauma), (b) propensity score method used (with sufficient information to allow reproducibility) and (c) a description of the detailed and justified inclusion of all relevant confounding co-variables. For propensity score matched articles, in addition to the above, we measured (a) the propensity score matching algorithm used [greedy matching (including nearest neighbor matching, +/- caliper distance) vs. optimal, with or without replacement], (b) balance of measured co-variables (using an absolute standardized difference <10%) and (c) the use of appropriate, matched statistical tests. A narrative (descriptive) synthesis was conducted. Our outcome measures included: (1) the proportion of studies adequately reporting propensity score generation, allowing for reproducibility, (2) the types of propensity score methodology utilized, (3) the appropriate use of propensity-score matching and reporting of covariate balance assessment, and (4) the use of appropriate statistical tests.

Results

A total of 700 potentially relevant citations were screened for retrieval from MEDLINE, EMBASE and Web of

Science. One hundred and eighteen duplicates were excluded. Five hundred and fifty-seven citations were excluded after scanning the titles and/or abstracts because they did not meet our inclusion criteria (Fig. 1). A total of 125 citations were retrieved for detailed evaluation of the full text studies. After excluding 54 studies, we were left with 71 studies.

Study characteristics

Of the 71 studies that used propensity score methods, there were 42 torso trauma studies, 14 emergency general surgery papers and 15 surgical critical care papers in total that used propensity score methods. Forty-three (61%) of these studies specifically used propensity score matching.

Included studies

The included studies were analyzed for report type, analysis, and matching.

Outcomes assessment

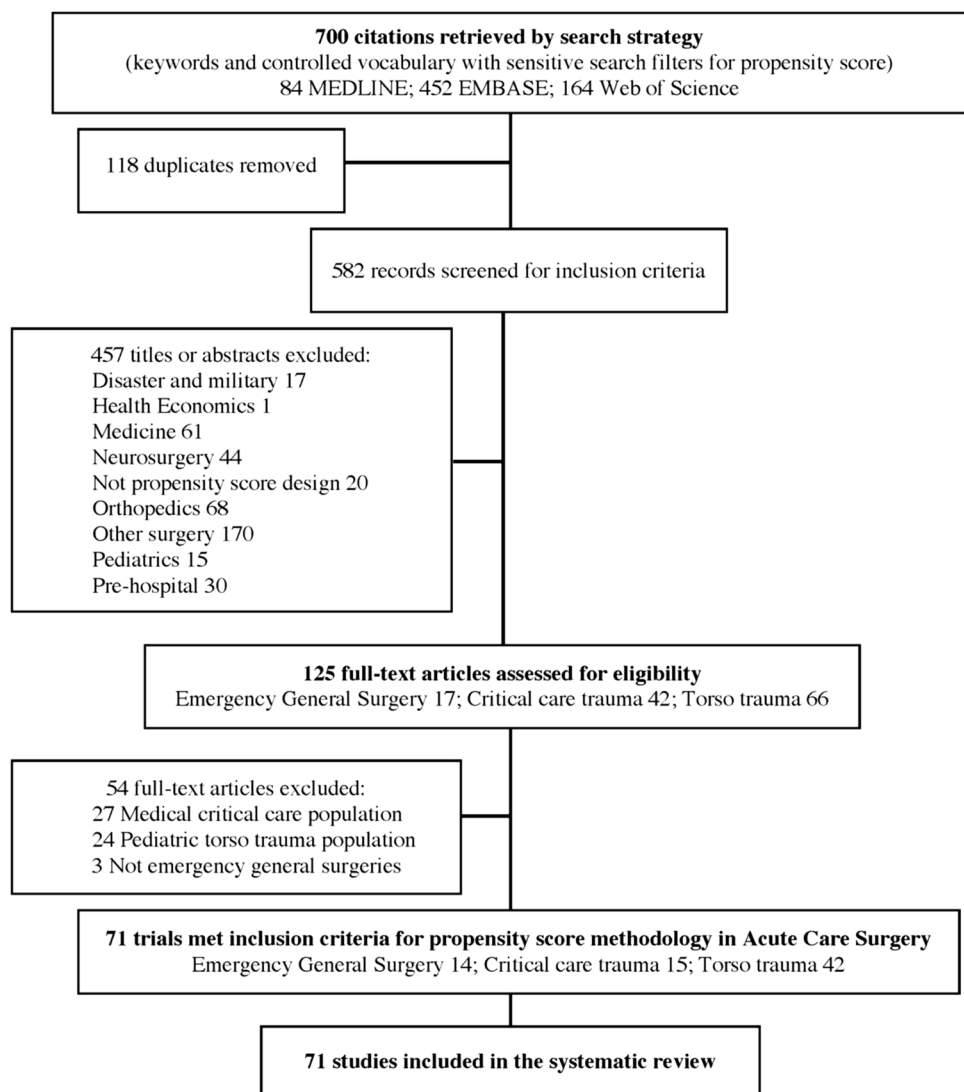
Propensity score reporting

Overall, of all the papers using propensity score methods, we found that over half of the papers ($n = 49$, 69%) were deemed adequate in their description of propensity score methods, allowing for reproducibility. The highest proportion of adequate, reproducible studies was found in surgical critical care (13/15, 87%) followed by trauma (27/42, 64%) and emergency general surgery (9/14, 64%). A total of 61/71 papers (86%) included some rationale for their included confounding co-variables in their propensity score model.

Type of propensity score methodology utilized

We found that the most common type of propensity score methods used in the ACS literature was propensity score matching (46/71, 65%). This was followed by adjustment (8/71, 11%), stratification (6/71, 8%), and weighting (2/71, 3%). There were 4% of papers (3/71) that described the use of computer software to generate their propensity score without further specification. One paper described the use of a cumulative logit model, while ten papers overall (14%) were not clear on the type of propensity score methods used.

Fig. 1 Flow diagram in accordance with the preferred reporting items of systematic reviews and meta-analyses (PRISMA) statement



Propensity score matching and assessment of covariate balance

A total of 46 studies overall (65%) described propensity score matching as the method of choice. Of these 46 studies, 36 (78%) matched studies were judged to have an adequate description of the method or algorithm used for matching. Nearest neighbor matching was the most common form of greedy matching (14/46, 30%), followed by the use of caliper matching in 28%. Greedy matching alone was described in two studies (4%), which lacked further details allowing for reproducibility. Forty one percent (19/46) of overall studies measured balance in their covariates after matching. Thirteen percent (6/46) of balanced studies used a standard difference of <10% to achieve balance. Nine studies (13% overall) measured the predictive ability of their propensity score model by

measuring the area under the receiver operating curve or a concordance statistic (c-statistic).

Appropriate statistical testing

Seventy six percent (35/46) of studies used appropriate, paired statistical tests to estimate treatment effect after matching. The most common errors included using Chi-squared tests, unpaired *t* tests, and Wilcoxon rank sum tests in the matched sample, the log-rank test to compare Kaplan–Meier survival curves in the matched sample, using standard Cox and logistic regression thereby failing to account for the matched nature of the sample. Overall, only 18/46 (39%) of studies reported an assessment of balance using standard differences and used the appropriate paired statistical tests for their models (Tables 1, 2, 3).

Table 1 Summary of eligible studies with overall assessment for adequacy. Studies are listed by year, with multiple studies per author per year listed chronologically.

Report characteristic	Year	Propensity score method clearly stated and reproducible?	Detail and justification of covariates?	If PS matching used: 1. Matching algorithm described? 2. Balance of measured co-variables? 3. Use of paired statistical tests?	Overall adequacy
Torso trauma (<i>N</i> =42)					
Aljuhani	2015	PS matching; non-reproducible	Yes	1. No 2. No 3. No	Inadequate
Allen	2015	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Beydoun	2014	Covariate adjustment; non-reproducible	Yes	N/A	Inadequate
Carlson	2013	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Cook	2015	PS matching; non-reproducible	No	1. No 2. No 3. No	Inadequate
Cudnik	2009	Unclear; non-reproducible	Yes	N/A	Inadequate
Ditillo	2014	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Dubose	2008	Unclear; non-reproducible	Yes	N/A	Inadequate
Ellis	2009	Weighting and stratification	Yes	N/A	Adequate
Ferraris	2010	Unclear; non-reproducible	Ref. only	N/A	Inadequate
Ferraris	2012	PS matching	Yes	1. Yes (ref.) 2. Yes 3. Yes	Adequate
Ferraris	2012	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Gage	2013	Unclear; non-reproducible	Yes	N/A	Inadequate
Garwe	2010	Stratification	Yes	N/A	Adequate
Gomes	2011	Covariate adjustment	Yes	N/A	Adequate
Harris	2012	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Inaba	2010	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Inaba	2010	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Innerhofer	2012	PS matching	Yes	1. Yes (ref.) 2. Yes 3. Yes	Adequate
Jones	2013	PS matching; non-reproducible	Yes	1. No 2. No 3. Yes	Inadequate
Joseph	2014	PS matching; non-reproducible	No	1. No 2. No 3. Yes	Inadequate
Joseph	2015	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Katsura	2013	Unclear; non-reproducible	No	N/A	Inadequate
Kent	2008	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Kothari	2012	PS matching; non-reproducible	Yes	1. No 2. No 3. Unclear	Inadequate
Lai	2014	PS matching; non-reproducible	Yes	1. No 2. No 3. No	Inadequate
MacKenzie	2006	Weighting	Yes	N/A	Adequate
Branco	2014	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Newgard	2007	PS matching	Yes	1. Yes (ref.) 2. No 3. No	Adequate
Norii	2015	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Orlando	2012	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Paryavi	2014	Unclear; non-reproducible	Yes	N/A	Inadequate
Perel	2014	Stratification	Yes	N/A	Adequate
Rhee	2015	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Sambasivan	2011	Unclear; non-reproducible	Yes	N/A	Inadequate
Tessier	2013	PS matching	No	1. No 2. No 3. No	Inadequate
Thompson	2014	PS matching	Yes	1. Yes 2. No 3. Unclear	Adequate
Tinetti	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Valle	2014	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Walther	2014	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Yeguiayan	2012	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Zarzaur	2011	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Emergency general surgery (<i>N</i> =14)					
Biancari	2014	Unclear; non-reproducible	No	N/A	Inadequate
DeMestral	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate

Table 1 (continued)

Report characteristic	Year	Propensity score method clearly stated and reproducible?	Detail and justification of covariates?	If PS matching used: 1. Matching algorithm described? 2. Balance of measured co-variables? 3. Use of paired statistical tests?	Overall adequacy
Gale	2015	PS matching	No	1. No 2. No 3. Unclear	Inadequate
Haas	2012	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Hemmila	2010	PS matching	No	1. Yes 2. Yes 3. Unclear	Adequate
Ingraham	2010	Stratification; non-reproducible	No	N/A	Inadequate
Kelly	2014	Unclear; non-reproducible	Yes	Unclear	Inadequate
Kuwabara	2011	Unclear; non-reproducible	No	N/A	Inadequate
Lee	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Moazzez	2013	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Smith	2012	Stratification	Yes	N/A	Adequate
Speicher	2012	PS matching	Yes	1. Yes 2. No 3. Yes	Adequate
Turan	2010	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Zafar	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Critical care trauma (<i>N</i> =15)					
Arabi	2013	Covariate adjustment	Yes	N/A	Adequate
Barmparas	2011	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Fletcher	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Griesdale	2008	Covariate adjustment	Yes	N/A	Adequate
Hyde	2015	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Iapichino	2010	PS matching (based on ref.)	Yes	1. No 2. No 3. Unclear	Inadequate
Joseph	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Kamps	2011	PS matching	Yes	1. Minimal 2. No 3. Unclear	Inadequate
Lee	2014	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Michalia	2012	Covariate adjustment; non-reproducible	Yes	N/A	Adequate
Neal	2009	Covariate adjustment	Yes	N/A	Adequate
O'Keefe	2012	Stratification	Yes	N/A	Adequate
Rubano	2013	PS matching	Yes	1. Yes 2. Yes 3. Yes	Adequate
Sowards	2014	Covariate adjustment	Yes	N/A	Adequate
Walkey	2011	Covariate adjustment	Yes	N/A	Adequate

Table 2 Characteristics of propensity score model

	Total articles	Matching	Adjustment	Stratification	Weighting	Total adequate	Total inadequate: (propensity score matched only)	Total inad- equately: (all others)
Torso Trauma	42	29	2	3	2	27 (64%)	7	8
Emergency General Surgery	14	9	0	2	0	9 (64%)	1	4
Critical Care Trauma	15	8	6	1	0	13 (87%)	2	0
Total	71	46	8	6	2	49 (69%)	10 (22%)	12 (48%)
Primary outcome								
Binary	49							
Continuous	13							
Time to event	7							
Other	2							

Table 3 Characteristics of propensity score matched analyses ($n = 46$)

Methodological characteristic	N (%)
Method of matching	
Greedy	5 (11)
Optimal	2 (5)
Nearest neighbor	14 (30)
Radius	0
Kernel	0
Mahalanobis metric	0
Caliper/caliper at 0.2 logit of PS	13 (28)
Digit matching (5/other)	3 (5 digits)
Balance ratio	
1:1	11 (26)
1:2	0
1:3	2 (5)
Replacement	
Yes	1 (2)
No	10 (23)
Not detailed	22 (51)
Covariate balance discussed	
Yes	16 (37)
No	27 (63)
#Appropriate evaluation of covariate balance	
Yes	10 (23)
No	6 (14)
AUC or c-statistic measured	
5 (12)	
*Paired analysis using matched dataset	
Yes	16 (37)
No	22 (51)
Not detailed	5 (12)

*Appropriate statistical testing of treatment effect/exposure: Paired t test, Wilcoxon signed rank test, McNemar test for correlated binary proportions, Kaplan Meier survival curves, Cox proportional hazards stratified on matched pairs, Cox proportional hazards with robust standard errors account for clustering in matched pairs, Conditional logistic regression using GEE, Agresti and Min method

Inappropriate: Chi-squared tests, unpaired t tests, and Wilcoxon rank sum tests in the matched sample, the log-rank test to compare Kaplan–Meier survival curves in the matched sample, using standard Cox regression, standard logistic regression

#Appropriate statistical testing of distribution of baseline characteristics: standardized differences

Review of methodology

Fundamentals of reporting an adequate propensity score

The propensity score is the probability (from 0 to 1) of a patient receiving a treatment based on observed baseline co-variates. In a recent propensity score analysis by Inoue *et al.* on the use of Resuscitative Endovascular Balloon Occlusion of the Aorta (REBOA) to attenuate severe

torso hemorrhage, the authors estimated the probability of receiving REBOA conditional on the measured and justified covariates that may lead to confounding (such as age, systolic blood pressure, intra-abdominal or retroperitoneal hemorrhage) [12]. This was done to strengthen the credibility of the estimated association between treatment assignment (REBOA) and the outcome (time to in-hospital mortality). This is analogous to random assignment in gold-standard randomized control trials, which balances co-variates by randomly assigning subjects to control or interventional groups. The only systematic “difference” between the randomized treatment groups would be exposure to treatment. Therefore, in a subset of patients with the same propensity score, the distribution of measured baseline co-variates between groups being compared will be similar with treatment exposure being the only difference. In order to generate the actual propensity score, a logistic regression model is usually fit, in which an indicator variable denoting treatment status is regressed on baseline characteristics. These baseline characteristics are often chosen as those independent variables that affect outcome [13]. After the propensity score is generated, matched pairs are formed, which can be accomplished in different ways, with strengths and weaknesses of each model.

Propensity score matching has several variations. A commonly employed variation in propensity score matching is the “nearest neighbor” technique. This allows, based on the score, for an exposed participant to be matched to the nearest control, randomly, if more than one control is the closest. There is no difference in minimum or maximum distance that is allowable with the nearest neighbor technique. Nonetheless, distance minimum can be measured and specified by assigning a caliper distance between nearest neighbors. There is no gold standard that exists for an adequate caliper distance however previous work by Austin *et al.* indicates that a caliper distance of 0.2 of the standard deviation of the logit of the propensity score is ideal in many settings [14]. Greedy and optimal matching are also commonly described algorithms for matching in the literature, with nearest neighbor and caliper matching being examples of greedy matching. Overall, despite controversy in the literature, it appears that both greedy and optimal matching have similar performance with one matching algorithm not superior to another [15]. The ratio of controls to exposed participants is also a consideration in propensity score matching. The usual ratio seen is 1:1 but is by no means restricted to this, with increasing numbers of controls to exposed participants leading to more use of data and slightly better precision but requires more complex statistical analysis. Matching has been the recommended method for pairing control and treated participants in order to attempt to provide unbiased estimation of the treatment effect [16].

Stratification is a method of propensity score analysis that includes categorizing participants within strata, which are often based on the quintiles of the propensity score. This method also needs to assess whether balance has been achieved among baseline co-variables within each of the propensity score strata. Stratification, while employed in the propensity score literature, does not reduce systematic bias as effectively as matching does, however increasing the number of strata can also improve this. The third propensity score method is covariate adjustment, which uses the propensity score as a co-variate in the model. Using this approach, the outcome variable, such as mortality, is regressed on an indicator variable denoting treatment status and the estimated propensity score. This method can also result in biased estimates of odds ratios and hazard ratios [17, 18]. Inverse probability of treatment weighting is a method based on estimating a weight for each subject that is equal to the reciprocal of the probability of receiving the treatment that the subject actually received. In other words, treatment subjects are weighted by the inverse of the propensity score while control subjects are weighted by the reciprocal of (1-propensity score). Weighting the sample using these weights results in a treatment selection not confounded by measured baseline covariates [19]. Some prior studies have demonstrated that within the comparison of different propensity score methods, propensity score matching eliminates a greater proportion of baseline differences of all methods. Propensity score matching and inverse probability of treatment weighting eliminate a greater proportion of baseline differences between treated and untreated subjects compared to other methods, reducing bias [4, 5].

Inoue *et al.* proceeded with propensity score matching as their approach with 1:1 pair matching within a specified caliper distance, meaning one control patient was matched to one exposed patient, subject to the constraint that their propensity score differed by less than the caliper amount [12]. Matching with or without replacement is another principle of propensity score methods, which should be specified, especially when using matching methodology. Matching with replacement means that a control participant can be matched to multiple exposed participants, resulting in a larger number of matched pairs. However, since this generates some pairs that are no longer independent (as they are sharing the same control participant), appropriate statistical tests must be used to account for this loss of independence between pairs. If the propensity score model has been correctly specified, then measured baseline covariates will be balanced between treated and untreated patients. Inoue *et al.* chose to measure balance by assessing the absolute standardized difference of variables, a recommended method to assess balance [20]. Balance is how we determine when the propensity score model has been adequately

specified. Estimation of treatment effect is next. Once you have paired matches of patients that are exposed or not to treatment, and you are satisfied that your propensity score model is appropriate, you can proceed with estimating your treatment effect with binary, continuous or time-to-event outcomes. Inoue *et al.* chose time to in-hospital mortality as their time-to-event outcome. In order to accomplish this accurately, statistical tests are used to determine treatment effect by direct comparisons of exposed and control groups. These statistical tests, however, must take into consideration the matched nature of the paired groups. Therefore, the paired samples *t* test would be used for continuous outcomes, McNemar test for dichotomous (binary) outcomes and Cox-proportional hazard for matched time-to-event outcomes. These represent the fundamental principles of propensity score methods.

While not a focus of this paper, the treatment of missing data has gained in importance increasingly throughout the medical literature. Biased results may be easily generated if included co-variables have a high rate of missing data, especially if these data are discarded. Variables can be classified as missing or unknown and can be matched based on this. Multiple imputation is also commonly used where missing variables are imputed, or ascribed a new value, based on the average of a completed data set. The average may be that of the treatment effect based on the propensity scores of a complete data set, or it may be that of the propensity score itself [21].

Assessing for model adequacy

The fundamental principle of assessing the adequacy of your propensity score model is by demonstrating that there is balance in measured baseline covariates between control and treated groups. There are two main ways to do this: (1) statistically significant differences between groups can be measured like a standard Table 1 of any randomized control trial. However, this approach is controversial, and its use has been discouraged or (2) measure balance between groups using the standardized difference. A commonly accepted cutoff for such a difference is <0.1 . If a difference of >0.1 is encountered, the model is deemed inadequate and more variables must be added to the existing model or the model must be modified through the inclusion of interactions or non-linear terms. Many studies report an area under the curve or concordance statistic (c-statistic) of the propensity score model, which measures its discriminatory power. This, however, does not give any consideration as to whether an important confounding variable has been omitted from the model or whether balance has been achieved [20]. Therefore the c-statistic is not used to determine adequacy. In fact, there is no association between any goodness-of-fit model and the ability of a propensity

score model to accurately balance prognostically important co-variables.

Discussion

This systematic review of propensity score methods is the first in the ACS literature. Our search identified 71 studies that used propensity score methods. We examined all adult acute care surgery patients in retrospective or prospective non-interventional observational trials that use propensity score methods over the last 10 years. In this review of propensity score matching in the ACS literature, we have found that while many researchers are indeed using matching after generating their propensity score, the matching methods are poorly described, with little assessment of balance of baseline covariates. Slightly more than a third of propensity score papers used appropriate paired statistical tests for the final estimate of the treatment effect after matching. This is not surprising given the increased usage of propensity score methods in the surgical literature, but the absence of formal reporting standards or guidelines in trauma or acute care surgery overall. Propensity score methods were first introduced in 1983 by Rosenbaum and Rubin [22], however, in contrast to other study types, no formal reporting guidelines exist in the literature for its use. The International Committee of Medical Journal Editors (ICMJE) has developed recommendations to review best practice and ethical standards in the conduct and reporting of research and other material published in medical journals. The purpose of these recommendations is to produce clear, reproducible and unbiased medical journal studies. The Equator Network (<http://equator-network.org>) follows their acronym (Enhancing the QUALity and Transparency Of health Research) in providing reporting guidelines used for manuscript submission to journals both in medicine and surgery. The editors of *the Journal of Trauma and Acute Care Surgery* recommend the use of the EQUATOR reporting guidelines when preparing or critically appraising manuscripts, however EQUATOR, to date, does not include reporting guidelines on propensity score methods.

By reviewing the ACS literature, we have found that while there is great variability in the use of propensity scores, many authors are indeed making efforts to report on propensity score methods and using matching as the main way to use the score, once generated. We have broken down the results into (1) adequacy of description of propensity score reporting, (2) type of propensity score method used, (3) in propensity score matching, checking for covariate balance and (4) the appropriate use of paired statistical tests for subsequent analyses. We found that the most common manner to estimate the propensity score is by using a logistic regression model with exposure to the treatment as the

dependent variable. Most researchers are also using matching as the principle way to analyze their data, especially in the trauma literature.

In order to plan and analyze a study using propensity score methods, we recommend that several steps be followed. Keep in mind that the observational study should be designed specifically and a priori for this kind of analysis. Early consultation with your statistician to formulate a statistical plan should be established prior to the analysis being conducted. This will allow the researcher to measure a maximal number of confounders in order to obtain an unbiased estimation of treatment effect. To further refine reporting guidelines developed in other specialties [11], we have developed the following recommendations: (1) plan your study as a propensity study a priori to identify as many confounders as possible to include. Researchers should have a clear justification for why such confounders are included or excluded. (2) Build your propensity score model by including all your confounders. The model most typically used is considered non-parsimonious in that it includes, once again, as many confounders as possible. This score can be estimated from baseline data, most commonly using logistic regression, which regresses treatment status onto baseline characteristics. This may be thus described as “non-parsimonious logistic regression” as many confounders and variables that affect outcome are included as possible in this model to generate your propensity score. Care should be taken to avoid the inclusion of variables that only affect treatment-selection, and that have no independent effect on the outcome. (3) Match subjects on the propensity scores you have estimated and fully describe the matching process. Report the degree to which baseline covariate balance was achieved in the matched sample. It is important to remember that matching does not imply precise matching, but rather matched subjects have similar propensity scores based on baseline covariates drawn from the same distribution, instead of two randomly selected subjects. Thus outcomes will be correlated within pairs [22]. We recommend matching or weighting over stratification and adjustment. With matching, we recommend that the matching process be described such that the results may be duplicated. When matching, ideally the mean difference between treated and control participants should be minimal. Nearest neighbor, with or without caliper distance, were the most commonly reported methods of matching in the trauma and acute care surgery literature. When matching using caliper distance, we suggest using a caliper width of 0.2 of the standard deviation of the logit of the propensity score. Additional considerations include the ratio of control to treatment participants. These can be 1:1, known as paired matching or one can choose many controls for each treatment participant (X:1). Variable to one matching may indeed use a larger data set and possibly reduce bias but the analysis is more

statistically cumbersome [14, 23]. (4) Check the balance to see if your matching has been successful for measured covariates after propensity score application. This step is fundamental in assessing for propensity score model adequacy. If baseline covariates are not balanced in your treated and control groups after propensity score application, especially if these co-variables are prognostically important for your outcomes, this will introduce bias into your estimates of treatment effect. A common way to check for balance between co-variables is to compute the standardized difference between co-variables. The method used for calculating standardized differences will depend on the matching model used, with a typical cutoff of for a standardized difference of <0.1 indicating a negligible difference between covariates. (5) Analyze your outcomes for your chosen endpoints (treatment effect or other) and ensure that your follow up statistical tests are appropriate specifically for matched pairs analysis. While some researchers assessed the c-statistic or area under the curve to measure the discriminatory power of the propensity score model, this is not necessary for assessing adequacy. A model may have excellent discriminatory power, yet still have important confounding variables excluded from analysis.

Limitations of propensity score methodology and our systematic review

Propensity score methodologies, such as matching, are tools that infer causality but it is still no replacement for the gold standard of randomization. Matching will only control for the differences on observed variables; however, if unobserved variables are not accounted for, bias may still be introduced [24]. Propensity score matching is not a substitute for randomization, which balances observed and unobserved variables. Overlap must exist between groups on the propensity score to make a strong support for causality. It will not be useful as a method if those patients with a high propensity score are treated and those with a low score are not. It is also important to be mindful of the types of variables included in the propensity score model, as some variables are important in predicting the exposure to treatment, while others are important in predicting outcomes and some will do both. This review is limited because we chose to exclude other forms of trauma such as non-multisystem trauma leading to orthopedic or neurosurgical injuries only.

Conclusions

While many studies in the ACS literature employ propensity score methodology, flaws exist in the reporting and implementation of these methods. Statistical practice in the

trauma and acute care surgery literature would be improved through the adoption of formal reporting guidelines for the use of propensity score methods.

Compliance with ethical standards

Conflict of interest Victoria McCredie, Peter Austin and Tanya L. Zakrisson declare that they have no conflict of interest.

References

1. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res.* 2011;46(3):399–424.
2. McMurry TL, Hu Y, Blackstone EH, Kozower BD. Propensity scores: methods, considerations, and applications in the. *Journal of Thoracic Cardiovascular Surgery.* *J Thorac Cardiovasc Surg.* 2015;150(1):14–9.
3. Luo Z, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research: pitfalls and prospects. *Med Care Res Rev.* 2010;67(5):528–54.
4. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Mak.* 2009;29(6):661–77.
5. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med.* 2008;27(12):2037–49.
6. Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. *Circ Cardiovasc Qual Outcomes.* 2008;1(1):62–7.
7. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg.* 2007;134(5):1128–35.
8. Higgins JPT, Green S. *Cochrane handbook for systematic reviews of interventions.* New York: Wiley; 2011.
9. Panig N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLoS One.* 2013;8(12):e83138.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* 2009;339:b2535.
11. Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive Care Med.* 2010;36(12):1993–2003.
12. Inoue J, Shiraishi A, Yoshiyuki A, Haruta K, Matsui H, Otomo Y. Resuscitative endovascular balloon occlusion of the aorta might be dangerous in patients with severe torso trauma: a propensity score analysis. *J Trauma Acute Care Surg.* 2016;80(4):559–67.
13. 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.* 2007;26(4):734–53.
14. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat.* 2011;10(2):150–61.
15. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med.* 2014;33(6):1057–69.

16. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J.* 2009;51(1):171–84.
17. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med.* 2007;26(16):3078–94.
18. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med.* 2014;33(7):1242–58.
19. Morgan SL, Todd JL. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociol Methodol.* 2008;38:231–81.
20. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.* 2009;28(25):3083–107.
21. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat Methods Med Res.* 2016;25(1):188–204.
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.
23. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 2):69–80.
24. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Stat Med.* 2005;24(10):1563–78.