Resuscitation 125 (2018) 66–69

Contents lists available at ScienceDirect

Resuscitation

journal homepage: www.elsevier.com/locate/resuscitation

Commentary and concepts

Propensity scores – A brief introduction for resuscitation researchers☆

Lars W. Andersen[a,b,*], Tobias Kurth[c]

[a] Research Center for Emergency Medicine, Department of Clinical Medicine, Aarhus University Hospital, Aarhus, Denmark
[b] Center for Resuscitation Science, Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[c] Institute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany

A R T I C L E I N F O

Keywords:
Propensity scores
Methods
Epidemiology
Cardiac arrest

A B S T R A C T

Propensity scores are commonly used in observational research. This article provides a brief introduction to propensity scores aimed for resuscitation researchers. We discuss the concept and calculation of the propensity score and how it can be used to adjust for confounding via regression adjustment, stratification, weighting or matching. The interpretation of these method is briefly discussed and the advantages and limitations of propensity scores are presented. Lastly, we provide some practical recommendations for the presentation of studies using propensity scores.

Introduction

Observational research plays an important role in resuscitation research, both as a method to provide preliminary data for randomized trials and to try to answer causal questions that are not amenable to randomized trials for ethical or practical reasons. Observational studies are prone to a number of biases that may result in invalid results such as confounding, information bias, and selection bias [1].

Confounding is the confusion of effects that happens when the main effect of exposure on an outcome is influenced by extraneous factors that meets a set of criteria and by doing so leads to the detection of a spurious association (or no association when one exists) between an exposure and an outcome. Confounding arises if extraneous factors exists that are directly or indirectly a cause (or a risk factor) for the exposure and for the outcome. By definition, such factors cannot be a consequence of the exposure or a mediator of the effect of the exposure on the outcome [1,2].

Several methods have been developed to account or adjust for potential confounding such as restriction, stratification, regression, weighting, and matching as well as methods that combine various aspects of these techniques. One of these methods is the propensity score which has gained increased popularity over the last decades. Some recent examples from the resuscitation literature includes Goto et al. comparing chest compression strategies in children with out-of-hospital cardiac arrest [3], Hamilton et al. assessing the relationship between prehospital physician involvement and survival after out-of-hospital cardiac arrest [4], and Sutton et al. examining the association between physiologic monitoring of cardiopulmonary quality during adult cardiac arrest and outcomes [5].

The aim of this brief introduction is to provide resuscitation researchers with an overview of the propensity score methodology. For a more comprehensive coverage of the topic, the reader is referred to other sources [6–9].

Propensity scores

Concept and calculation

The propensity score was first introduced more than three decades ago as an alternative method to estimate effects in observational research [10,11]. The propensity score is the conditional probability (i.e. a value between 0 and 1) of assignment to an exposure of interest given observed characteristics [6,9,11]. The propensity score is usually estimated using logistic regression with the exposure of interest as the dependent variable (i.e. the exposure serves as the "outcome" of this model). The usual assumptions and modeling restraints of the logistic regression model still applies [12]. In theory, only variables that lead to true confounding (i.e. variables related to both the exposure and the outcome) need to be included in the model to produce unbiased results but these variables can be difficult to precisely characterize and identify [9]. Inclusion of variables associated with the outcome but not the exposure might result in decreased variance (i.e. narrower confidence intervals) [13,14]. Including many potential covariates to control for confounding (e.g. all potentially relevant variables collected in a

☆ A Spanish translated version of the abstract of this article appears as Appendix in the final online version at https://doi.org/10.1016/j.resuscitation.2018.02.003.

* Corresponding author at: Research Center for Emergency Medicine, Department of Clinical Medicine, Aarhus University Hospital, Nørrebrogade 44, Building 1B, 1st floor, 8000 Aarhus C, Denmark.
E-mail address: lwandersen@clin.au.dk (L.W. Andersen).

https://doi.org/10.1016/j.resuscitation.2018.02.003
Received 21 December 2017; Received in revised form 26 January 2018; Accepted 5 February 2018
0300-9572/

registry) might result in less precise estimates but this is likely of relatively less importance with very large sample sizes [14,15]. In some instances, it might also be relevant to include interaction terms especially if subgroups exist for which the exposure-outcome association differs [16]. Importantly, as with other adjustment tools to control for confounding, mediators of the effect between the exposure and the outcome should not be included. For example, in a study examining the association between automated external defibrillators and survival, it would be inappropriate to adjust for time to defibrillation since time to defibrillation might be a consequence of automated external defibrillator use and also influence improved survival. In general, characteristics that occur after the exposure should not be included.

Missing data is relatively common within observational resuscitation research and should be explicitly addressed including when using propensity scores. The topic of missing data is beyond the scope of the current manuscript and guidance can be found elsewhere [17]. When missing data is common, complete case analysis (i.e. only including patients with complete data) could give biased results and should generally be avoided [17].

Use of the propensity score

Once the propensity score is calculated from the logistic regression model, it can be used in four different ways to adjust for confounding: regression adjustment, stratification, weighting, or matching [9]. Although all methods are valid, matching and weighting seems to provide the most unbiased results [6,9]. However, the most appropriate method depend on the research question and the population to which this research question applies [18].

With regression adjustment, the propensity score is included in a regression model as an independent variable along with the exposure of interest. This method is relatively straight forward and computationally easy but requires that the relationship between the propensity score and the outcomes is modelled correctly (e.g. linear, polynomial, categorical).

With stratification, the propensity score is divided into several strata (often quintiles) [19]. The exposure-outcome association is then estimated in each strata, and either reported separately, or more commonly combined using appropriate statistics.

The propensity score can be used for weighting in several ways, the most commonly used being inverse probability of treatment weighting. With this approach a pseudo-cohort is created where covariates leading to confounding are balanced between exposed and unexposed patients. The pseudo-cohort is created by assigning each patient a weight; for the exposed, this weight is 1 divided by the propensity score and for the unexposed it is 1 divided by 1 minus the propensity score. Additional details on this approach has been provided elsewhere [20].

With matching, patients are matched based on the propensity score to create comparable groups. Matching requires several considerations including the unit of matching (propensity score vs. logit of the propensity score), the number of matches (e.g. 1:1, 1:2), the type of matching (e.g. optimal, nearest neighbor), the caliber of matching, whether to replace the unexposed, and the order of the matching. A detailed discussion of these considerations is beyond the scope of the current manuscript and has been provided elsewhere [9,21,22]. Once matching is performed, the baseline characteristics in the matched groups should be compared using standardized differences [9,21,22]. Formulas for standardized differences have been provided elsewhere [9]. Generally a standardized difference less than 0.1 is considered negligible [9,23]. The use of p-values for this comparison is discouraged since p-values can be misleading in both small (non-significant p-values might be due to a lack of power) and large (significant p-values might not represent a relevant difference) sample sizes. It is not recommended to assess the discriminatory power of the models with the c-statistic [24]. When comparing outcomes, it is generally recommended that statistical analyses should account for the matching [9].

It is important to note that no propensity score method, including matching, adjusts for unmeasured confounders. In the case of matching, the exposed and unexposed are matched on included characteristics which does not in any way guarantee matching on other variables not included in the propensity score model. The use of the term "pseudo-randomized" for propensity score matched cohorts is therefore discouraged.

Interpretation of results

Results from observational studies using propensity score methods should be interpreted with the same caution as other observational studies. Propensity score methods only adjust for measured covariates leading to confounding and the reader should therefore be cognizant about the risk of biased results due to unmeasured or residual confounding. Unmeasured confounding arises when not enough covariates are available to fully control for confounding. Residual confounding is confounding that persists if measured covariates are not measured correctly (i.e., misclassification, measurement error), or misspecified (e.g. dichotomizing a continuous confounder when the confounding relationship is more complex) [25–27].

The various propensity score methods provide estimates of effects that might differ [18]. For example, when using matching, the estimated effect reflects the effect in the matched cohort which in many studies will be different from the overall cohort of interest. For example, Kitamura et al. assessed the association between public access defibrillation and survival at 1 month with a favorable neurologic outcome using propensity score matching [28]. 43,776 patients were included in their cohort of interest, however only 8442 patients were included in the propensity score matched cohort. These patients had important differences in patient characteristics compared to the entire cohort (for example, near 100% bystander cardiopulmonary resuscitation [CPR] compared to only 56% in the overall cohort) [28]. If the effect of bystander defibrillation differs according to bystander CPR (i.e. effect modification), the effect of bystander defibrillation found in the matched cohort might not represent the effect of bystander defibrillation in the population of interest. For a more detailed example on this concept, see Kurth et al. [18].

Comparison to regression

Although propensity scores are gaining substantial popularity, it is important to note that they theoretically are no better or worse than traditional regression methods under similar set-ups and assumptions. Both methodologies adjust for measured and included covariates to control confounding and assume that no unmeasured or residual confounding persists to estimate causal effects [9]. The methods also tend to produce similar results [9,29,30].

There are some potential advantages to propensity score matching as compared to traditional regression [9]. First, as noted by Rosenbaum and Rubin, matching allows "… researchers to appreciate immediately the equivalence of treatment [exposed] and control [unexposed] groups …" [11] which might be preferable to the "black box" of regression where the details of the analysis is often unclear to most investigators. This also allows for an examination of the overlap in baseline characteristics between the groups. In extreme situations where baseline characteristics (i.e. the propensity score) only overlap between a very small proportion of the study cohort, there might be important limitations to the results [9]. Second, matching (and weighting) allows for a separation of the "design" and the analysis stage of the study allowing the investigator to calculate the propensity score and perform the matching, including potential changes to the propensity score model to optimize balancing of covariates, without analyzing or seeing any results related to the outcome(s) [9]. Third, propensity score methods are preferable in situation where the outcome is rare but the exposure is common and there is a large number of potentially covariates to adjust