Journal of Clinical Epidemiology

Graphic report of the results from propensity score method analyses

Ian Shrier[a,*], Menglan Pang[b], Robert W. Platt[b]

[a]*Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, 3755 Cote Ste-Catherine Road, Montreal H3T 1E2, Canada*
[b]*Department of Epidemiology, Biostatistics and Occupational Health, Purvis Hall McGill University, 1020 Pine Ave. West, Montreal, Quebec H3A 1A2, Canada*

Abstract

Objectives: To increase transparency in studies reporting propensity scores by using graphical methods that clearly illustrate (1) the number of participant exclusions that occur as a consequence of the analytic strategy and (2) whether treatment effects are constant or heterogeneous across propensity scores.

Study Design and Setting: We applied graphical methods to a real-world pharmacoepidemiologic study that evaluated the effect of initiating statin medication on the 1-year all-cause mortality post-myocardial infarction. We propose graphical methods to show the consequences of trimming and matching on the exclusion of participants from the analysis. We also propose the use of meta-analytical forest plots to show the magnitude of effect heterogeneity.

Results: A density plot with vertical lines demonstrated the proportion of subjects excluded because of trimming. A frequency plot with horizontal lines demonstrated the proportion of subjects excluded because of matching. An augmented forest plot illustrates the amount of effect heterogeneity present in the data.

Conclusion: Our proposed techniques present additional and useful information that helps readers understand the sample that is analyzed with propensity score methods and whether effect heterogeneity is present. © 2017 Elsevier Inc. All rights reserved.

Keywords: Propensity score; Trimming; Effect heterogeneity; Meta-analysis; Density plot; Frequency matched plot

1. Introduction

Observational studies usually encounter confounding when estimating the effect of treatment [1]. One method used to control for confounding is the propensity score [2]. The propensity score is defined as the probability of receiving the treatment conditional on covariates [3]. It is commonly estimated using logistic regression and is considered as a summary score for the included covariates. Subjects with identical propensity scores have, on average, the same prognosis and can be treated as exchangeable, if the key assumptions hold for positivity, consistency, no unmeasured confounding, and correct model specification [4].

There are several ways in which the propensity score can be used to estimate causal effects including stratification, matching, regression adjustment, and inverse probability weighting [5–7]. Using propensity score methods is only appropriate if the probability of receiving any level of treatment (conditional on the covariates) is greater than zero for each participant in the analysis [4,8]. Practically, one way this can be verified is if there is no subject with a propensity score extremely close to 0 or 1 and if the propensity score distributions of the two treatment groups overlap throughout their full range. Sometimes the full data set includes some participants with very low (or high) propensity scores, or for which there is no participant in the other treatment group with the same propensity score. In these contexts, investigators will use one of two common approaches to restrict the population analyzed so that the assumption is true on the analyzed population. First, investigators may "trim" (exclude) those participants who have extreme propensity scores from the study population, as recommended by Stürmer et al. [9,10]. Trimming can be performed based on the regions of nonoverlap of the estimated propensity score, percentiles of the estimated

* Corresponding author. Tel.: +1-514-340-8222 ext 4244; fax: +1-514-340-7564.

E-mail address: ian.shrier@mcgill.ca (I. Shrier).

<table>
<tbody>
<tr><td colspan="4">**What is new?**</td></tr>
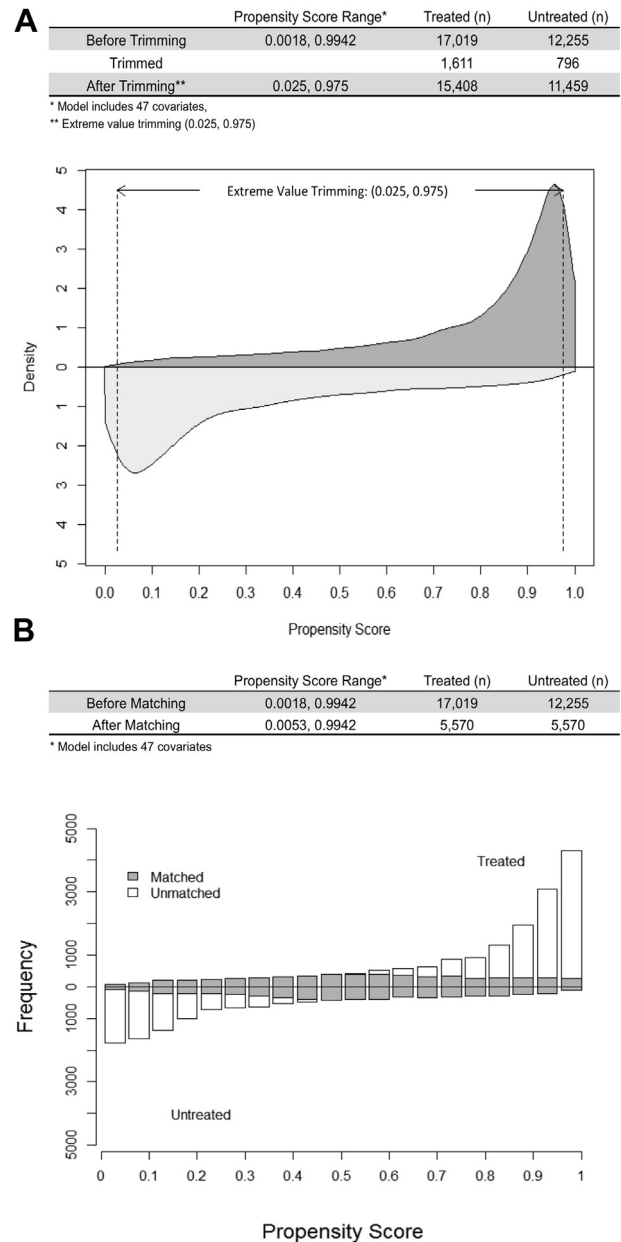</tbody>
</table>

What is new?

Key findings

- Graphical methods may be helpful in understanding how propensity score methods alter the population under study.

- An augmented density plot for propensity score trimming, or an augmented frequency plot for propensity score matching, increases the transparency of the analysis when propensity score methods are used.

- Methods analogous to meta-analytical techniques are helpful to assess if treatment effects are heterogeneous across strengths of indication for treatment and therefore if a single propensity score estimate is appropriate.

to difficulties in interpreting the standard mean (SD) table of comparisons. Fig. 1 in our article provides a general overview of how the study population is altered in terms of propensity score distribution before and after matching or trimming on propensity score. If significantly altered, this may lead to challenges in interpretation. Without appropriate transparency, readers and decision makers may make incorrect inferences based on the results provided.

A	Propensity Score Range*	Treated (n)	Untreated (n)
Before Trimming	0.0018, 0.9942	17,019	12,255
Trimmed		1,611	796
After Trimming**	0.025, 0.975	15,408	11,459

* Model includes 47 covariates,
** Extreme value trimming (0.025, 0.975)

B	Propensity Score Range*	Treated (n)	Untreated (n)
Before Matching	0.0018, 0.9942	17,019	12,255
After Matching	0.0053, 0.9942	5,570	5,570

* Model includes 47 covariates

Fig. 1. (A) A density plot illustrates the participants included in the analysis before and after trimming, with associated information about propensity score methods. (B) A frequency-matched plot is used to best illustrate the participants included in the analysis before and after matching. Both analyses use data from a study of statin and 1-year all-cause mortality post-myocardial infarction [14].

propensity score, or prespecified extreme values. Although trimming may ensure overlap of propensity scores, the distribution of propensity scores between the two treatment groups will generally be very different and adjustment as described previously is still required. Second, investigators may match participants in the treatment group to one or more participants in the untreated group with the same propensity score (perhaps many-to-many). Matching on the exact propensity score is usually not feasible, and nearest neighbor matching with a certain caliper is recommended [11,12]. With a relatively narrow caliper, participants that have no comparator with respect to the propensity score will be eliminated from the population.

Both trimming and matching result in exclusions from the sample. One recommendation within the STROBE statement is to enhance transparency of observational studies through a participant flow diagram [13]. This diagram should illustrate how many participants were originally approached, how many had complete follow-up, and reasons for excluding participants [13]. Although it is possible to include a line in the participant flow diagram indicating the additional exclusions due to trimming and/or matching, the actual exclusions due to matching occur at specific propensity scores that cannot be easily conveyed with text. Furthermore, these exclusions may change the population being analyzed considerably, such that both trimming and matching may change the parameters of interest compared with the original sample population. Current standard practice includes the presentation of the mean and standard deviation of each baseline covariate before and after matching. This is important because it provides direct evidence for imbalance of potential confounders. However, when there are many variables in the propensity score, there are very likely to be meaningful differences between groups for different variables. This leads