ELSEVIER

Journal of Clinical Epidemiology ■ (2014) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine

Zhongheng Zhang*, Hongying Ni, Xiao Xu

Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, No. 351, Mingyue Street, Jinhua, Zhejiang, 321000, P.R. China

Accepted 21 February 2014; Published online xxxx

Abstract

Background and Objective: Propensity score (PS) analysis has been increasingly used in critical care medicine; however, its validation has not been systematically investigated. The present study aimed to compare effect sizes in PS-based observational studies vs. randomized controlled trials (RCTs) (or meta-analysis of RCTs).

Methods: Critical care observational studies using PS were systematically searched in PubMed from inception to April 2013. Identified PS-based studies were matched to one or more RCTs in terms of population, intervention, comparison, and outcome. The effect sizes of experimental treatments were compared for PS-based studies vs. RCTs (or meta-analysis of RCTs) with sign test. Furthermore, ratio of odds ratio (ROR) was calculated from the interaction term of treatment × study type in a logistic regression model. A ROR < 1 indicates greater benefit for experimental treatment in RCTs compared with PS-based studies. RORs of each comparison were pooled by using meta-analytic approach with random-effects model.

Results: A total of 20 PS-based studies were identified and matched to RCTs. Twelve of the 20 comparisons showed greater beneficial effect for experimental treatment in RCTs than that in PS-based studies (sign test $P = 0.503$). The difference was statistically significant in four comparisons. ROR can be calculated from 13 comparisons, of which four showed significantly greater beneficial effect for experimental treatment in RCTs. The pooled ROR was 0.71 (95% CI: 0.63, 0.79; $P = 0.002$), suggesting that RCTs (or meta-analysis of RCTs) were more likely to report beneficial effect for the experimental treatment than PS-based studies. The result remained unchanged in sensitivity analysis and meta-regression.

Conclusion: In critical care literature, PS-based observational study is likely to report less beneficial effect of experimental treatment compared with RCTs (or meta-analysis of RCTs).

Keywords: Propensity score; Randomized controlled trial; Critical care; Effect size; Ratio of odds ratio; Observational study

1. Introduction

Well-designed and properly conducted randomized controlled trial (RCT) is one of the most important sources of evidence for clinical decision-making. Randomization will balance both measured and unmeasured variables between treated and untreated subjects. RCT can provide causal association between intervention and outcome, which is the key for clinicians to understand the underlying mechanisms for a pathologic condition. However, such experimental studies are often not feasible because of economical and ethical constraints [1]. Thus, clinical evidence is often shaped by observational studies, in which however the treatment effect is often confounded by many measured and unmeasured factors. Many techniques have been developed to control these confounding factors, including stratification, matching, and multivariable regression analysis [2].

Propensity score (PS) analysis was developed in the 1980s and has been increasingly used in biomedical field [3–6]. It is defined as the conditional probability of receiving a treatment or exposure given a series of predefined covariates [7]. With conventional matching or stratification, only few covariates can be taken into account, whereas the PS technique is able to incorporate all measured confounding factors and assigned each subject a score based on the probability that one will receive treatment. PS can be used for adjustment, matching, weighting, and stratification [8]. Critical care studies are especially subjected to bias because a long list of baseline characteristics cannot be easily balanced, and there is large number of interventions

Conflict of interest: None.

* Corresponding author. Tel./fax: +86-579-82552629.
E-mail address: zh_zhang1984@hotmail.com (Z. Zhang).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.02.018

What is new?

- The present study demonstrates that PS-based observational study is likely to report less beneficial effect of experimental treatment compared with RCTs in the area of critical care medicine.

other than experimental treatment being conducted in intensive care unit (ICU). Thus, it is of crucial importance to control confounding factors in observational studies, particularly when administrative data are used for analysis [9]. Thus, PS has found its way into the field of critical care medicine, and the number of publications involving PS has increased exponentially in recent years [10]. However, the validity of PS has long been debated, and it is unknown whether the result obtained by using PS is comparable with that obtained by RCTs. Thus, the present study aimed to compare the treatment effect for experimental intervention in PS-based observational studies vs. RCTs (or meta-analysis of RCTs) in critical care medicine.

2. Methods

2.1. Study selection

Observational studies using PS in the field of critical care medicine were identified by searching PubMed from inception to April 2013. There was no language restriction. Searching strategies consisted terms related to critical care and PS and mortality: (((((critically ill[Title/Abstract]) OR critical care[Title/Abstract]) OR intensive care[Title/Abstract]) OR ICU[Title/Abstract]) AND propensity score[Title/Abstract]) AND mortality[Title/Abstract]. Studies were potentially eligible if they (1) were related to critical care medicine; (2) used PS as a technique to adjust for pretreatment variables; (3) involved human subject; and (4) reported mortality as an end point. Exclusion criteria were (1) studies that investigated risk factors for mortality (not treatment effect); (2) intervention was not in the field of critical care medicine, for instance, studies in cardiothoracic surgery were excluded; (3) PS studies that cannot be matched to an RCT; details for matching were described in the following; and (4) the reported effect size could not be matched to that in corresponding RCTs; for instance, PS-based study reported hazards ratio (HR) but the RCT reported odds ratio (OR).

Each observational study using PS analysis was matched to one or more RCTs. Although the matching process was inherently subjective, every effort was made to match a PS-based study with RCTs in terms of population, intervention, comparison, and outcome (PICO) [11]. If more than one RCTs were identified, the effect sizes were combined by using meta-analytic approach with random-effects model [12].

Data on the following aspects were abstracted from the PS-based observational studies: name of the first author, year of publication, sample size, the number of RCTs being matched, study design (eg, prospective or retrospective), techniques for the using of PS (eg, matching, weighting, adjustment, and stratification), the number of covariates used to obtain PS, type of the effect size, and topic area. If data were not explicitly reported, we would try to contact the contributing author for detailed information.

2.2. Statistical analysis

The reported effect sizes included OR, relative risk (RR), and HR. If a study did not report OR and it used PS-matching technique, OR was calculated in the matched cohort. The effect sizes were compared between PS study and RCT (or meta-analysis of RCTs) by using binomial (sign) test to see whether one type of study design was more likely to report beneficial effect than the other. For studies that reported the number of survivors and nonsurvivors, we established a logistic regression model to calculate the relative effect size (ratio of OR, ROR) and associated 95% confidence interval (CI) [13]. The model was based on the equation: logit (π) = $\beta_0|\beta_1 I_t|\beta_2 I_{tp}|\beta_3 I_p$, where π is the probability that an event is observed; I_t, I_{tp} and I_p are variables denoting the effect of treatment (I_t = 1 in treatment subjects, 0 otherwise), the treatment—PS interaction (I_{tp} = 1 in treated subjects in studies using PS, 0 otherwise), and the effect of PS design (Ip = 1 in studies using PS, 0 otherwise); β_s were parameters of the logistic regression model. ROR can be obtained from the estimated β_2. ROR was obtained for each pair of matched PS study and RCTs. An ROR $<$ 1 indicates there is a greater benefit for experimental treatment in RCTs (or meta-analysis of RCTs) compared with the PS-based study; conversely, an ROR $>$ 1 suggests that there is a greater benefit for experimental treatment in PS-based study. Finally, RORs of matched pairs were combined by using meta-analytic approach with random-effects model. We predefined that the RCT with the largest sample size was the ''gold standard'' for the real treatment effect, and sensitivity analysis was performed by restricting to RCTs with the largest sample size. All statistical analyses were performed using the software StataSE 11.2 (StataCorp, College Station, TX, USA). Two-tailed P $<$ 0.05 was considered to be statistically significant.

3. Results

3.1. Study selection and characteristics

Fig. 1 shows the flow chart of study selection. Our initial search identified 161 potential studies. Among them, 109 studies were excluded because they were not critical care studies, investigating risk factors or not involving human subjects. The remaining 52 studies using PS were matched