ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1076−1084

Journal of Clinical Epidemiology

Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery

O. Kuss[a,*], T. Legler[a], J. Börgermann[b]

[a]*Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, University of Halle-Wittenberg, 06097 Halle (Saale), Germany*
[b]*Department of Thoracic and Cardiovascular Surgery, Heart and Diabetes Center North Rhine-Westphalia, Ruhr-University Bochum, 32545 Bad Oeynhausen, Germany*

Accepted 20 January 2011

Abstract

Objective: Analyses comparing randomized to nonrandomized clinical trials suffer from the fact that the study populations are usually different. We aimed for a comparison of randomized clinical trials (RCTs) and propensity score (PS) analyses in similar populations.

Study Design and Setting: In a systematic review, we ''meta-matched'' RCTs and PS analyses that compared the off- and the on-pump technique in coronary artery bypass grafting. ''Meta-confounders'' were summarized in a ''meta-propensity score'' and were used for ''meta-matching.'' We compared treatment effects between RCTs and PS analyses for 10 previously defined binary clinical outcomes in this ''meta-matched'' population as differences in ''meta-odds ratios.''

Results: For all clinical outcomes, the estimated differences in ''meta-odds ratios'' were below an absolute value of 0.15, all confidence intervals included the null.

Conclusions: In our example, treatment effects of off-pump versus on-pump surgery from RCTs and PS analyses were very similar in a ''meta-matched'' population of studies, indicating that only a small remaining bias is present in PS analyses.

Keywords: Randomized controlled trial; Propensity score; External validity; Internal validity; Off-pump coronary artery bypass

1. Introduction

It is commonly agreed that randomized clinical trials (RCTs) are the gold standard for treatment evaluation. However, RCTs have also often been criticized for limited external validity, which results from the enrollment of highly selected patients groups [1]. Study patients tend to be younger and healthier than the average patient [2,3]. Nonrandomized clinical trials (non-RCTs) or observational studies are a possible alternative for assessing treatment effects because those are expected to have larger external validity. Their obvious principal disadvantage is limited internal validity as nonrandomized treatment allocation might bias treatment comparisons because of confounding [4]. A wealth of methods to adjust for this confounding have been proposed [5], the most recent (although developed as early as in the 1980s) being the technique of propensity score (PS) analysis [6], which is expected to have statistical advantages as compared with the standard methods of confounder adjustment [7−9].

Numerous investigations have been conducted to assess if treatment effects from RCTs differ systematically from those of non-RCT studies. Most of these investigations have already been collected in systematic reviews [2,10−13]. Evidence from these reviews is still inconclusive, results from non-RCT studies differ sometimes, but not always, and not in a predictable direction from the results of RCTs.

A simple consequence of the limited *external* validity of RCTs is, however, the limited *internal* validity of all systematic comparisons of RCTs and non-RCTs: If RCTs are conducted in highly selected populations, but non-RCTs in the general population, differences between both study types are not necessarily because of the missing randomization. They might also arise from the different study populations involved.

Ideally, we would like to conduct a ''meta-randomized'' trial to systematically compare RCTs and non-RCTs. That is, investigators willing to conduct a study on a specific

Funding: There was no external funding for this study.

* Corresponding author. Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, Martin-Luther-University of Halle-Wittenberg, Magdeburger Str. 8, 06097 Halle (Saale), Germany. Tel.: +345-557-3582; fax: +345-557-3580.

E-mail address: oliver.kuss@medizin.uni-halle.de (O. Kuss).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.01.005

What is new?

- Most analyses comparing randomized to non-randomized clinical trials suffer from the fact that the study populations are different.
- By a ''meta-propensity score analysis,'' we generated a ''meta-matched'' sample of randomized clinical trials (RCTs) and propensity score (PS) analyses that assessed the off-pump technique in coronary artery bypass grafting.
- Treatment effects in the ''meta-matched'' sample of RCTs and PS analyses were similar, indicating that only a small remaining bias is present in PS analyses.

clinical question would be randomly selected to perform an RCT or a non-RCT. Proceeding this way, all characteristics of the investigator's setting and patients would distribute evenly on the group of RCTs and non-RCTs, thus eliminating all kinds of ''meta-confounding'' and enabling a causal statement on the effect of randomization. Obviously, conducting such a study would be difficult and maybe even unethical, as we should not force investigators to conduct non-RCTs if they are willing to do an RCT. It is interesting to note that similar trials have already been performed, not on a study, but on a patient level [14–17]. In these trials, usually called doubly randomized preference trials, patients have been randomized in a first stage to (1) a group, where the actual treatment would be subsequently randomized in a second stage or (2) to a group where they choose the treatment they prefer.

If ''meta-randomized'' trials are difficult, one is still able, however, to conduct a ''meta-non-randomized'' trial to compare RCTs and non-RCTs. As PS analyses are the most valid methods (at least in statistical terms [7–9]) for analyzing non-RCTs, we conducted a ''meta-propensity score analysis'' to judge the differences between RCTs and non-RCTs. That is, we ''meta-matched'' RCTs and non-RCTs for important and available ''meta-confounders,'' where the latter were summarized in a ''meta-propensity score.'' We then compared treatment effects between RCTs and non-RCTs in this ''meta-matched'' population.

As a clinical example, we use the comparison of off-pump versus on-pump technique in coronary artery bypass grafting (CABG), one of the most hotly debated and polarizing issues in cardiac surgery [18]. The example was chosen because the authors knew from previous work [19] that enough PS analyses and RCTs would be available for a meaningful analysis. This clinical question has also some public health relevance. For example, in Germany, 49.788 (isolated) bypass surgeries were performed in 2007, 10.1% of those used the off-pump technique [20].

2. Methods

2.1. Study selection—selection of non-RCTs

In the group of non-RCTs, we restricted ourselves to PS analyses. We included all PS analyses comparing off- and on-pump CABG from our recent systematic review [19], details on study selection, and search strategy are given there. As the study search for this review was performed in February 2006, we again performed the described search in October 2006. PS analyses were included in the analysis presented here if they gave descriptive information on the PS study publication (e.g., average age, proportion of males etc., factors we will subsequently refer to as ''meta-confounders'') and at least one of the 10 short-term binary clinical outcomes death, stroke, myocardial infarction (MI), atrial fibrillation, acute renal failure, inotropic support, RBC transfusion, wound infection, reoperation for bleeding or, intra-aortic balloon pump support.

2.2. Study selection—selection of RCTs

Randomized trials were retrieved by including all publications that were referred to in the five largest systematic reviews [21–25]. As the literature search for the most recent review [24] included studies only up to February 2006, we additionally performed a MEDLINE search for the keywords ''randomized'' and ''off-pump'' in October 2006. All RCT publications were gathered in full text and read independently by two reviewers (O.K., T.L.), blinded for the results of the previous reviews. Data were abstracted into a self-developed case record form, which had been tested in a small (five studies) pilot study. Results from the two reviewers were entered in an MS ACCESS database. Disagreements between reviewers were found by automatic comparisons and resolved by (in this hierarchical order) finding a consensus, discussion with a third reviewer (J.B.), or referring to the previous systematic reviews. RCTs were included if they gave descriptive information on the RCT study publication and at least one of the binary clinical outcomes mentioned above. Double publications were removed, but we used all additional information on study populations and outcomes from the removed publications.

2.3. Extracted information

From each of the retrieved studies (PS analyses and RCTs), we extracted the observation period, location of study, number of involved centers, number of treated patients, and percentage of conversions to on-pump. We further extracted any information that described the study populations, that is, all preoperative prognostic and risk factors provided.

To describe the estimated treatment effects from each study, we separately extracted the original fourfold tables for each of the 10 binary outcomes from the RCTs and