Journal of Clinical Epidemiology 68 (2015) 1232–1240

Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review

Hervé Laborde-Castérot[a,b], Nelly Agrinier[a,c], Nathalie Thilly[a,c,*]

[a]*Lorraine University, Paris-Descartes University, EA 4360 Apemac, Avenue de la forêt de Haye, 54500 Vandoeuvre-lès-Nancy, France*
[b]*Université Paris 13, Sorbonne Paris Cité, UFR SMBH, 1 rue de Chablis, 93017, Bobigny, France*
[c]*Clinical Epidemiology and Evaluation, CIC-EC CIE6 Inserm, University Hospital of Nancy, Allée du Morvan, 54500 Vandoeuvre-lès-Nancy, France*

Accepted 2 April 2015; Published online 8 April 2015

Abstract

Objectives: Propensity score (PS) and instrumental variable (IV) are analytical techniques used to adjust for confounding in observational research. More and more, they seem to be used simultaneously in studies evaluating health interventions. The present review aimed to analyze the agreement between PS and IV results in medical research published to date.

Study Design and Setting: Review of all published observational studies that evaluated a clinical intervention using simultaneously PS and IV analyses, as identified in MEDLINE and Web of Science.

Results: Thirty-seven studies, most of them published during the previous 5 years, reported 55 comparisons between results from PS and IV analyses. There was a slight/fair agreement between the methods [Cohen's kappa coefficient = 0.21 (95% confidence interval: 0.00, 0.41)]. In 23 cases (42%), results were nonsignificant for one method and significant for the other, and IV analysis results were nonsignificant in most situations (87%).

Conclusion: Discrepancies are frequent between PS and IV analyses and can be interpreted in various ways. This suggests that researchers should carefully consider their analytical choices, and readers should be cautious when interpreting results, until further studies clarify the respective roles of the two methods in observational comparative effectiveness research.

Keywords: Instrumental variable; Propensity score; Confounding by indication; Observational studies; Comparative effectiveness research; Statistical methods

1. Introduction

Evidence-based medicine has conferred on randomized controlled trials (RCTs) a high level of evidence concerning the results on efficacy of clinical interventions. RCTs minimize bias and control confounding and are therefore considered the gold standard of design validity [1]. However, efficacy does not necessarily mean effectiveness. RCTs are generally conducted under ideal conditions, among highly selected patients followed by hyperspecialized physicians, and often fail to demonstrate the generalizability of their results in a real-world setting [2]. Moreover, it is not always possible, for practical or ethical reasons, to carry out an RCT [2]. Thus, complementary, or alternative approaches when RCTs are not possible, is needed to evaluate the effectiveness of clinical interventions. Well-designed observational studies may be useful to evaluate real-world usage patterns and the effects of clinical interventions [2,3]. However, they are prone to bias, particularly to confounding by indication: assignment to intervention does not occur by chance but depends on patient characteristics that can influence the effect of the intervention on the outcome [4]. In other words, the apparent effectiveness of an intervention may be explained by preintervention differences in risk factors between patients who received the

Conflict of interest: None.
Funding: None.

* Corresponding author. Service d'Epidémiologie et Evaluation cliniques—CHU de Nancy, Hôpitaux de Brabois. Allée du Morvan - 54500 Vandœuvre-lès-Nancy. Tél.: (33) 3-83-85-21-63; fax: (33) 3-83-85-12-05.
E-mail address: n.thilly@chu-nancy.fr (N. Thilly).

http://dx.doi.org/10.1016/j.jclinepi.2015.04.003

What is new?

Key findings

- More and more observational studies simultaneously use propensity score (PS) and instrumental variable (IV) approaches to evaluate the same intervention, often leading to nonconcordant results that may be difficult to interpret.
- Discrepancies between results of the two methods can be explained by an expected difference in the control of confounding by indication, but also by theoretical differences between methods or by unintended consequences of inappropriate use.

What is the implication and what should change now?

- Researchers should be aware of the impact on results of the analytical technique chosen (PS or IV) and its appropriate use in observational studies, and readers should be cautious in the interpretation of results.
- Further studies are needed to investigate the agreement between PS and IV analyses in particular settings and precise the indications for each method.

intervention and those who did not. Consequently, analytical techniques are needed to address the problem of confounding in observational data [5]. Because of its ability to control for numerous potentially confounding factors, propensity score (PS) analysis, proposed by Rubin and Rosenbaum [6,7], has been increasingly used in this context for 15 years. PS represents the likelihood of a patient being assigned to an intervention on the basis of his or her preintervention characteristics. It may be applied in different ways: matching, stratification, adjustment, and weighting [8]. As compared with traditional regression models, the number of potential confounders considered in the analysis is not limited. Thus, PS methods theoretically increase comparability between groups by creating pseudorandomization of all possible measured confounders. However, even if PS method is able to reduce bias due to all measured confounders, it fails to limit bias due to unmeasured or unknown confounders [7].

To address this limit, instrumental variable (IV) analysis [9,10], widely used in economy during the last decades [11], has recently emerged in the field of clinical research and is increasingly used. This technique compares patient groups according to an IV—also named instrument—which is randomly distributed, rather than comparing patients with respect to the actual intervention received. A critical step is to find an appropriate instrument that should meet three requirements: (1) to be associated with the intervention (relevancy assumption); (2) not to directly affect the outcome of interest, but only indirectly affect it through the intervention assignment (exclusion restriction); and (3) to be independent of confounders [9,10]. Theoretically, if properly implemented, IV analysis differs from PS analysis in that it also aims to control for unmeasured or unknown confounders.

To date, the optimal approach (PS or IV) to adjust for confounders in observational studies has remained unclear, and researchers tend increasingly to use both methods—and compare results—to evaluate the effectiveness of an intervention. The way to use PS on one hand and IV on the other in medical research has been the focus of several reviews [8,12–17], but no publication has yet compared results obtained with both methods for the same analyses. These comparisons would help understand whether both methods lead to concordant results or not and interpret their respective findings. The purpose of this article was to systematically review the current medical literature in which the effects of interventions are estimated by both PS and IV analyses and to discuss the agreement between these two analytical methods.

2. Methods

2.1. Search strategy

A comprehensive search of the literature in MEDLINE and in the medical research part of Web of Science was performed to identify all published observational studies that evaluated a clinical intervention using both PS and IV analyses. We used the combination of terms "propensity" and "instrumental variable(s)," and limited the research to studies published up to March 31, 2014 (see Appendix at www.jclinepi.com, for search strategy). Additional publications have been identified by screening the references of the full-text articles selected.

2.2. Study selection

Records identified using the previously mentioned databases were independently screened by two authors on their titles and abstracts. Studies were eligible for the analysis if they satisfied the following selection criteria: (1) evaluation of an intervention using both PS and IV methods for the same analyses; (2) reporting of quantitative results for both methods, even if they were expressed differently, with a confidence interval (95% CI) or *P*-value for significance testing; (3) use of morbidity or mortality criteria as outcomes; and (4) publication in English. Records with insufficient description of the methodology, such as brief reports or congress abstracts, were not considered for review. If insufficient information on selection criteria was available in the abstract, the full-text article was considered. Finally, studies that met the criteria after a full-text assessment were included in the review.