Systematic review of propensity score matching in anesthesiology

Jack O. Wasey waseyj@chop.edu, Mohamed A. Rehman rehman@chop.edu
Children’s Hospital of Philadelphia, PA, USA

Background

Pediatric anesthesia is harder to study prospectively than in adults, and many things of interest are not suited to randomized controlled trials. Propensity score matching is widely used in retrospective studies to reduce potential confounding factors, but many are conducted or reported inadequately. This systematic review assesses the quality of published studies using propensity score matching, and compares their methods and reporting to best-practice.

The development and rationale for the propensity score model is important: the covariates used should be listed, a rationale for why those were chosen, and at least a minimal description of the model type, which is typically, but not necessarily, logistic regression. Calipers are often used to prevent poorly matched controls when all the good matches have been used up. Replacement is when controls are allowed to be replaced into the pool for matching again, thus allowing one control to match multiple cases.

Balance diagnostics should be provided, ideally beyond just comparing means.

Methods

Pubmed was searched for articles in the journals Pediatric Anesthesiology and Anesthesiology with publication dates between October 2014 and October 2016 inclusive, for the terms “propensity” or “matching.”

The search found 19 matching articles, of which 17 used propensity scores to match cases to controls. None of these articles met recommended reporting standards, being deficient in key elements of propensity model construction or balance diagnostics, or both.

Results

Propensity matching is the most popular matching method, but has major problems. The model used to generate propensity score is important and should be reported in detail, but often is not. Whichever matching technique is chosen, balance diagnostics should be performed, but this is rarely done properly. Without these things, the validity of matching, and therefore of whole studies must be questioned, because bias may have been increased. Contrary to popular belief, causal effects can be estimated from retrospective data, which is widely practised in non-medical disciplines, but results can only be useful when the matching is done well.

Conclusions

References

