Discussion: Confounding in Observational Studies

Frank E Harrell Jr

Department of Biostatistics
Vanderbilt University School of Medicine

NIH AHRQ Methodologic Challenges in CER

2 December 2010

Outline

Experimental vs. Observational Research

- What characterizes situations when causal inference from observational data can be trusted?
  - Prognostic factors well understood and collected
  - Rich, accurate, data purposefully collected
  - Treatment by indication well understood/characterized
  - Reproducible research
  - Good statistical analysis practice
  - Pre-specified analytic plan
- RCTs usually utilize inflexible *P*-value-based approaches
- Need quick launch pragmatic randomized trials
- RCTs vs. endlessly debated observational results

Current State of Observational Research

The majority of observational etiologic, treatment, and epidemiologic research is either

- wrong
- not reproducible
- overstated

Making Observational Research Rigorous

- Be objective; avoid confirmation bias
- Mask outcome data while formulating analysis
- Approximate RCTs in methodology
- Pre-filed statistical analysis plan
  - document & justify exceptions
- May be best to quit if key covariates are not collected

Rubin [2007, 2008, 2010]

Objective Determination of Adequacy of Covariates

Example: Connors et al. [1996] – Effectiveness of right heart catheterization

- Panel of 4 intensivists and 3 cardiologists specified variables deemed to relate to the decision to use a RHC
- All variables were available from the intensive prospective data collection
- 70 parameter propensity model
- Extensive sensitivity analysis for unmeasured confounders

Confounder Adjustment Methods

- Standard covariate adjustment
  - When no unassailable instrument and when no. of baseline covariates is, say, $<$ effective sample size/10, can't be beat
- Instrumental variables
- Propensity scores (PS)
  - a data reduction method
  - very useful if outcomes rare but all treatments commonly used
  - assumptions are easier to understand

Propensity Score Methods

- Liberal use of baseline variables
- Allow covariates to be nonlinear and sometimes to interact
- May be useful to use machine learning techniques to predict treatment received (Westreich et al. [2010])
- Matching and stratification are greatly overused
  - discards significant amount of data
  - arbitrary
  - residual confounding especially in outer tails (Lunceford and Davidian [2004])

PS Methods, *continued*

- Covariate adjustment using logit PS
  - expand using regression splines; don't assume linearity
- Limit analysis to covariate overlap regions
- Pre-specified "big prognostic players" also included separately
- Parsimony is not sought
  - significance tests for covariates in PS model inappropriate
- PS is a way of aggressively adjusting for **observable** covariates

Example Interpretation of PS Analyses

A 0.95 confidence interval for the hazard ratio for treatment A compared to treatment B is $[0.65, 0.83]$ adjusted for 40 baseline covariates in a PS and also for 5 pre-specified clinically important prognostic factors (which were also in PS).

Variables related to both treatment selection and outcome are assumed to be in the model or to be easily predicted from combinations of variables in the model.

Variables are assumed to have adequate measurement accuracy.

The confidence interval excludes 1.0 as long as an unmeasured confounder has an odds ratio ≤ 2 in predicting that a patient will get treatment A and a hazard ratio ≤ 1.7 in predicting time to clinical endpoint.

What is Evidence?

- Not a measure of non-randomness, surprise, or embarrassment (P-value)
- Confidence interval taking all uncertainties into account
  - including uncertainty about adequacy of available covariates
- Bayesian credible intervals
- Likelihood support intervals

Summary

- Observational research has many challenges and is easy to do poorly
- It must be made rigorous to be respected
- Propensity analysis is a useful tool in observational treatment comparisons
- Such analysis needs to be done carefully and comprehensively
- Should be accompanied by a sensitivity analysis, and limitations noted
- We need to be better educators about what evidence really means

There is nothing wrong with cancer research that a little less money wouldn't cure.

Nathan Mantel, NCI

A. F. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, W. A. Knaus, and The SUPPORT Investigators. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276:889–897, 1996.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*, 23:2937–2960, 2004.

D. B. Rubin. The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized studies. *Stat Med*, 26:20–36, 2007.

D. B. Rubin. For objective causal inference, design trumps analysis. *Ann Appl Stat*, 2(3):808–840, 2008.

D. B. Rubin. On the limitations of comparative effectiveness research. *Stat Med*, 29:1991–1995, 2010.

D. Westreich, J. Lessler, and M. J. Funk. Propensity score estimation: neural networks, support vector machines, decision

trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epi*, 63:826–833, 2010.