Combining inverse-probability weighting and multiple imputation to adjust for selection bias due to missing data in EHR-based research

Sebastien Haneuse

Harvard T.H. Chan School of Public Health
E-mail: shaneuse@hsph.harvard.edu

Abstract: Among the many potential threats to validity in EHR-based studies, selection bias due to missing data is prominent. Existing missing data methods, however, such as inverse-probability weighting (IPW) and multiple imputation (MI), typically fail to acknowledge the complexity of EHR data. To resolve this, Haneuse et al (2016) proposed to modularize the data provenance into a series of sub-mechanisms, each representing a clinical "decision". Based on this we develop a general and scalable framework for estimation and inference for regression models that permits the use of IPW and/or MI to tackle each of the sub-mechanisms in a single analysis. We refer to this as a "blended analysis strategy". We show that the proposed estimator is consistent and asymptotically Normal, and derive a consistent estimator of the asymptotic variance. Simulations show that naïve use of standard methods may result in bias; that the proposed estimators have good small-sample properties; and, that a bias-variance trade-off may manifest as researchers consider how to handle missing data. The proposed methods are illustrated with data from a multi-site EHR-based study of the effect of bariatric surgery on BMI.