Computational Statistics and Data Analysis 72 (2014) 219–226

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases

Jessica M. Franklin*, Sebastian Schneeweiss, Jennifer M. Polinski, Jeremy A. Rassen

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St., Suite 3030, Boston, MA 02120, USA

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form 30 August 2013
Accepted 18 October 2013
Available online 28 October 2013

Keywords:
Simulation
Pharmacoepidemiology
Propensity score
Variable selection

ABSTRACT

Longitudinal healthcare claim databases are frequently used for studying the comparative safety and effectiveness of medications, but results from these studies may be biased due to residual confounding. It is unclear whether methods for confounding adjustment that have been shown to perform well in small, simple nonrandomized studies are applicable to the large, complex pharmacoepidemiologic studies created from secondary healthcare data. Ordinary simulation approaches for evaluating the performance of statistical methods do not capture important features of healthcare claims. A statistical framework for creating replicated simulation datasets from an empirical cohort study in electronic healthcare claims data is developed and validated. The approach relies on resampling from the observed covariate and exposure data without modification in all simulated datasets to preserve the associations among these variables. Repeated outcomes are simulated using a true treatment effect of the investigator's choice and the baseline hazard function estimated from the empirical data. As an example, this framework is applied to a study of high versus low-intensity statin use and cardiovascular outcomes. Simulated data is based on real data drawn from Medicare Parts A and B linked with a prescription drug insurance claims database maintained by Caremark. Properties of the data simulated using this framework are compared with the empirical data on which the simulations were based. In addition, the simulated datasets are used to compare variable selection strategies for confounder adjustment via the propensity score, including high-dimensional approaches that could not be evaluated with ordinary simulation methods. The simulated datasets are found to closely resemble the observed complex data structure but have the advantage of an investigator-specified exposure effect.

1. Introduction

Longitudinal healthcare claims databases are frequently used for studying the comparative safety and effectiveness of medications. Administrative healthcare data generally provide a longitudinal record of medical services, procedures, diagnoses, and medications for large numbers of patients and therefore provide a rich data source for conducting pharmacoepidemiologic research. Compared with randomized trials, the data available in healthcare claims better represent the full spectrum of patients that are exposed to a drug and the processes of care in routine practice (Schneeweiss and Avorn,

* Correspondence to: 1620 Tremont St., Suite 3030, Boston, MA 02120, USA. Tel.: +1 617 278 0675; fax: +1 617 232 8602.
E-mail address: JMFranklin@partners.org (J.M. Franklin).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.018

2005; Strom and Carson, 1990). However, drug studies in claims data may suffer from bias due to residual confounding (Brookhart et al., 2010), and it is unclear whether methods for confounding adjustment that have been shown to perform well in small, simple nonrandomized studies are applicable to the cohort studies created from complex healthcare claims data.

Monte Carlo simulation can be used to evaluate the performance of causal inference methods, but ordinary simulation approaches do not capture important features of healthcare claims. For example, healthcare claims databases often have hundreds, or even thousands, of measured covariates with complex covariance structures. These covariates, either singly or in combination, may serve as proxies for unmeasured confounders and be effectively used to remove bias (Schneeweiss et al., 2009). Further, the complexity of real-world data extends beyond confounding; patients' follow-up time and censoring patterns are often associated with exposure and outcome via a path of underlying characteristics. These complexities cannot be replicated in fully synthetic simulated data, as they are generally not completely understood and vary by data source.

As an alternative to simulation, Vaughan et al. (2009) suggested creating "plasmode" datasets. A plasmode is a real dataset that is created from natural processes but has some aspect of the data-generating model known, for example, a "spike-in" experiment in microarray analysis of gene expression where a known amount of genome transcript is added to each sample. Merging this concept with simulation techniques has led to several studies of methods performance that use real observed data augmented with simulated data (Elobeid et al., 2009; Gadbury et al., 2008). Other approaches utilize fully simulated data, but create associations among variables to match estimated associations from observed data (Chao et al., 2010; Erenay et al., 2011; McClure et al., 2008; Rolka et al., 2005; Schmidt et al., 2009), including one approach specifically designed for simulating an entire healthcare claims database (Murray et al., 2011). However, due to the massive size and complexity of the data in this approach, generally only one dataset is created for each set of simulation parameters, and the relative contributions of bias and variance to estimation error cannot be judged. Furthermore, this data-generation process may produce spurious correlations among variables that are not present in the underlying empirical dataset.

In this paper, we outline a statistical and computational framework for creating replicated simulation datasets based on an empirical pharmacoepidemiologic cohort study in healthcare claims data. The objective of this work is to enable the evaluation of approaches to confounder adjustment in simulated data that preserve the complex features and information content of claims data but also have a known true treatment effect. As an example, we applied our framework to a study of high versus low-intensity statin use and cardiovascular outcomes. Simulated data was based on real data drawn from Medicare Parts A and B and eligibility files linked with Part D prescription drug insurance claims database maintained by Caremark. We compared properties of the data simulated using this framework with the empirical data on which the simulations were based. In addition, we used the simulated datasets to compare variable selection strategies for confounder adjustment via the propensity score (PS), including high-dimensional approaches that could not be evaluated with ordinary simulation methods since their performance depends on the information richness and complexity of the underlying empirical data source.

2. Methods

Our simulation approach relies on resampling from the observed covariate and exposure data without modification in all simulated datasets to preserve the empirical associations among these variables. Repeated outcomes are simulated using a true treatment effect of the investigator's choice and the baseline hazard function estimated from the empirical data (Fig. 1). R code and documentation for the simulation setup are available in the Web Appendix.

2.1. Construct the cohort

The first task in creating simulated datasets is to create the cohort on which the simulations will be based from the larger healthcare database. The specifics of the study design, including inclusion and exclusion criteria for the cohort, definitions of exposures and covariates, and determinations of follow-up and censoring for outcome events, are important in determining the performance of any statistical methods subsequently applied to the data. As these issues are not the focus of this paper, we refer the reader to the literature on the subject for specific information on these determinations. In general, we recommend a "new user design" with an active comparator (Ray, 2003; Schneeweiss, 2010), where two treatments with similar clinical indications are compared in patients initiating one treatment or the other with no history of use in the prior six months (or some other pre-specified period). Covariates (diagnoses, procedures, medications, and health system service use) are assessed in the period preceding initiation of treatment, and assessment of outcomes begins on or after the date of treatment initiation.

The result of this design is a dataset where each patient has information on exposure ($X = 1$ indicates initiating one treatment, $X = 0$ indicates initiating the reference treatment), presence of an outcome event (Y), and length of follow-up time (T). In addition, we assume that there is a large pool of potential covariates, C, that contains potentially hundreds or thousands of distinct codes for diagnoses, procedures, hospitalizations, medications and other health system service use in the period preceding treatment initiation (Schneeweiss et al., 2009). This dataset provides all of the information that we will use for constructing the simulated datasets.