Computers in Biology and Medicine 69 (2016) 61–70

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Refining adverse drug reaction signals by incorporating interaction variables identified using emergent pattern mining

Jenna M. Reps [a,*], Uwe Aickelin [a], Richard B. Hubbard [b]

[a] School of Computer Science, Jubilee Campus, University of Nottingham, NG8 1BB, United Kingdom
[b] School of Medicine, University of Nottingham, Nottingham NG5 1PB, United Kingdom

ARTICLE INFO

Article history:
Received 24 August 2015
Accepted 24 November 2015

Keywords:
Medical informatics
Signal refinement
Data mining
Observational data
Confounding
Emergent pattern mining

ABSTRACT

Purpose: To develop a framework for identifying and incorporating candidate confounding interaction terms into a regularised cox regression analysis to refine adverse drug reaction signals obtained via longitudinal observational data.

Methods: We considered six drug families that are commonly associated with myocardial infarction in observational healthcare data, but where the causal relationship ground truth is known (adverse drug reaction or not). We applied emergent pattern mining to find itemsets of drugs and medical events that are associated with the development of myocardial infarction. These are the candidate confounding interaction terms. We then implemented a cohort study design using regularised cox regression that incorporated and accounted for the candidate confounding interaction terms.

Results: The methodology was able to account for signals generated due to confounding and a cox regression with elastic net regularisation correctly ranking the drug families known to be true adverse drug reactions above those that are not. This was not the case without the inclusion of the candidate confounding interaction terms, where confounding leads to a non-adverse drug reaction being ranked highest.

Conclusions: The methodology is efficient, can identify high-order confounding interactions and does not require expert input to specify outcome specific confounders, so it can be applied for any outcome of interest to quickly refine its signals. The proposed method shows excellent potential to overcome some forms of confounding and therefore reduce the false positive rate for signal analysis using longitudinal data.

1. Introduction

Negative side effects of medication, termed adverse drug reactions (ADRs), are a serious burden to healthcare [1,2]. ADRs are estimated as the cause of 6.5% of UK hospitalisations [2] and a study investigating US death due to ADRs reported rates between 0.08 and 0.12 per 100,000 [3]. Studies have suggested that the rate of ADRs is increasing annually [4], motivating the improvement of methods for detecting them.

The process of detecting ADRs starts during clinical trials, however clinical trials often lack sufficient power to detect all ADRs for numerous reasons including time limitations, unrealistic conditions and a limited number of people being included [5]. It is then down to post-marketing surveillance to identify the remaining undiscovered ADRs. This involves three stages: signal detection (identifying associations between drugs and outcomes), signal refinement (prioritising/filtering spurious relationships) and signal evaluation (confirming causality after numerous sources of evidence). There has been a big focus towards developing signal detection methods, involving various forms of data such as spontaneous reporting systems [6], online data [7,8], chemical structures [9] and longitudinal observational data [10,11]. Unfortunately, all the data sources have their own limitations. Spontaneous reporting systems are historically the main source used for post-marketing analysis but often contain missing values, suffer from under- and over-reporting, and rely on people noticing ADRs [12]. Longitudinal observational data have recently been used to complement spontaneous reporting system data for extracting new drug safety information, and are an excellent potential source of information due to the quantity of observational data available and the number of variables recorded. If we could overcome existing issues, mainly confounding, that limit the use of observational data for causal inference then we may be able to aid the discovery of new ADRs.

* Corresponding author.
E-mail address: jenna.m.reps@gmail.com (J.M. Reps).

http://dx.doi.org/10.1016/j.compbiomed.2015.11.014
0010-4825/

We are often plagued with confounding when investigating potential causal relationships retrospectively in observational data [13] due to the data collection being non-random. When an association between an exposure and outcome is discovered in observational data, it may often be explained by the presence of confounding. A confounding variable is one that leads to distorted effect estimates between an exposure and outcome due to the confounder being associated with both the exposure and outcome. For a variable to be considered a confounder of an exposure and outcome relationship it must be a risk factor of the outcome, it must be associated with the exposure and it cannot lie within the causal pathway between the exposure and outcome.

Consider, for example, the situation where we wish to determine the relationship between a drug given to treat hypertension and myocardial infarction. If we naively look at the incidence of myocardial infarction within a year after treatment for patients given the drug and the incidence of myocardial infarction within a randomly chosen year for patients never given the drug, then we are likely to find that myocardial infarction is more common in those given the drug and conclude that the drug is associated with an increased incidence of myocardial infarction. However, our conclusion is likely explained by confounding, as patients given the drug (those with hypertension) are medically different from those who do not have hypertension. It is likely that some of the patients given the drug have a poor diet or are stressed. Poor diet and stress would have contributed to the hypertension but are also risk factors of myocardial infarction. Therefore poor diet and stress would be confounding factors. To correctly determine a relationship between an exposure and outcome it is important to account for confounding variables. Techniques such as risk adjustment, stratification, or equally distributing the confounding variables between the comparison groups are potential ways to reduce confounding [14].

Adjusting for confounders in observational data requires identifying the confounders. Although existing methods aim to address confounding, various studies have shown that existing signal generation methods developed for longitudinal observational data have a high false positive rate [15,16]. This is most likely due to difficulties identifying confounding variables in a data-driven way. Some studies have shown that including a large number of variables, such as drug indications, into drug safety methods can reduce confounding [17–19], but none of these methods included interactive terms. A medical illness is likely to be a result of multiple variables interacting. For example, cardiovascular disease is common in patients with a genetic predisposition such as familial hypercholesterolemia and based on lifestyle such as diet and exercise. Therefore, it is interactive terms between medical events or drugs that are most likely to correspond to confounding variables. However, when there are thousands of medical events and drugs, the number of possible interactions is very large. Existing data-driven methods for incorporating interactive terms into regression models include hierarchal lasso, which adds the interactions along with an interaction regularisation term [20], and methods utilising matrix factorisation [21]. However, these methods are likely to be highly inefficient when there are thousands of variables to consider (which is often the case for observational data). Instead, methods such as emergent pattern mining [22] that can efficient identify outcome specific associations, even when large numbers of variables are being considered, may be more suitable. A similar idea was used to successfully detect survival associate rules [23] based on cox regression and association rule mining. This shows that it is possible to reduce confounding by combining cox regression and association rule mining.

A suitable post-marketing framework that extracts knowledge from longitudinal observational data could be of the form displayed in Fig. 1. The first stage of the proposed framework is to apply an efficient large-scale signal generation method to find associations between exposures and outcomes. In the first step the method would efficiently search through all the exposure and outcome possibilities to find associated pairs. An example of a suitable signal generation method is the high dimensionality propensity score (HDPS) [24]. The HDPS works by developing a predictive model for taking the drug and then a matched cohort analysis is applied, where controls are selected based on having a high propensity for taking the drug (the predictive model predicts that they would have the drug). The HDPS can limit confounding by accounting for a large number of variables. Unfortunately, it is not without issues [25,26] and still often signals many false positives [15], this highlights the requirement of additional analysis that can reduce the false positive rate. The second step in the framework is the signal refinement, where complex confounding relationships are discovered and incorporated into a more detailed analysis. The output of the signal refinement is a small set of exposure–outcome pairs that are prioritised for signal evaluation. The final step would be to formally evaluate the remaining signals using a number of different data sources, as establishing a causal relationship requires an accumulation of evidence.

In this paper we focus on the signal refinement stage, as there are no data-driven methods to refine signals, but numerous signal generation and evaluation methods exist. The objective of this research is to develop a data-driven signal refinement methodology that can be applied after ADR signal generation using longitudinal observational data to filter and re-rank the signals by addressing complex confounding. We will test the data-driven methodology by analysing the relationship between numerous drugs and the outcome myocardial infarction (MI). We are exploring three goals:

1. Whether emergent pattern mining can be used to identify candidate interaction confounding covariates in a data-driven way.
2. Whether the inclusion of interaction confounding covariates into a regression analysis can reduce confounding and be used for data-driven ADR signal refinement.
3. Whether lasso and ridge regularisation are suitable techniques to enable the inclusion of a large number of potential interaction covariates.

2. Materials and methods

2.1. Materials

The longitudinal observational database used in this study is The Health Improvement Network (THIN) database (http://www.thin-uk.com. THIN contains complete medical records for patients registered at a participating general practice within the UK. At present approximately 6% of the UK general practices are participating, resulting in THIN containing data on over 4 million active patients. The validity of the THIN database for pharmacoepidemiology studies has been investigated [27] and it was shown that the data appear to be representative of the UK population.

THIN contains time-stamped entries of drugs that are prescribed and medical events that the general practices are made aware of. Prescribed drugs are recorded with a British National Formulary (BNF) code indicating the family of the drug prescribed. Medical events are recorded using the Read coding system. The Read codes used in this study to identify myocardial infarction (MI) are available in Supplementary data.

The drug families (represented by BNF codes) investigated in this study are presented in Table 1 along with the ground truth (the known relationship between each drug family and MI), the number of prescriptions eligible for inclusion in the study and number of MI