Journal of Biomedical Informatics 56 (2015) 356–368

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A supervised adverse drug reaction signalling framework imitating Bradford Hill's causality considerations

Jenna Marie Reps [a,*], Jonathan M. Garibaldi [a], Uwe Aickelin [a], Jack E. Gibson [b], Richard B. Hubbard [b]

[a] School of Computer Science, University of Nottingham, NG8 1BB, UK
[b] Division of Epidemiology and Public Health, University of Nottingham, UK

ARTICLE INFO

Article history:
Received 16 March 2015
Revised 8 June 2015
Accepted 15 June 2015
Available online 24 June 2015

Keywords:
Big data
Pharmacovigilance
Longitudinal observational data
Causal effects
Signal detection

ABSTRACT

Big longitudinal observational medical data potentially hold a wealth of information and have been recognised as potential sources for gaining new drug safety knowledge. Unfortunately there are many complexities and underlying issues when analysing longitudinal observational data. Due to these complexities, existing methods for large-scale detection of negative side effects using observational data all tend to have issues distinguishing between association and causality. New methods that can better discriminate causal and non-causal relationships need to be developed to fully utilise the data.

In this paper we propose using a set of causality considerations developed by the epidemiologist Bradford Hill as a basis for engineering features that enable the application of supervised learning for the problem of detecting negative side effects. The Bradford Hill considerations look at various perspectives of a drug and outcome relationship to determine whether it shows causal traits. We taught a classifier to find patterns within these perspectives and it learned to discriminate between association and causality. The novelty of this research is the combination of supervised learning and Bradford Hill's causality considerations to automate the Bradford Hill's causality assessment.

We evaluated the framework on a drug safety gold standard known as the observational medical outcomes partnership's non-specified association reference set. The methodology obtained excellent discrimination ability with area under the curves ranging between 0.792 and 0.940 (existing method optimal: 0.73) and a mean average precision of 0.640 (existing method optimal: 0.141). The proposed features can be calculated efficiently and be readily updated, making the framework suitable for big observational data.

1. Introduction

Side effects of prescription drugs, also known as adverse drug reactions (ADRs), occur unpredictably and present a major healthcare issue. It is possible that a generally healthy individual may take a prescription drug for a minor problem and end up with a potentially life threatening ADR. As a consequence, it is essential to monitor all marketed drugs and develop methods that are capable of identifying ADRs at the earliest possible point in time. The potential benefits of utilising longitudinal observational data for detecting (also known as signalling) ADRs have been highlighted [1]. However, unsupervised methods developed to signal ADRs using longitudinal observational data have been found to obtain high false positive rates consistently across data sources [2,3]. This is due to the complexities of observational data, such as missing data and confounding, making it difficult for the methods to distinguish between association and causality. Reference sets detailing known ADRs and non ADRs have been created to aid the development of ADR signalling methods for longitudinal data by enabling a fair evaluation of the methods' ADR signalling performances [4]. However, the creation of reference sets now presents the opportunity of generating labelled data and developing a supervised framework that can be applied to longitudinal observational data to signal ADRs. The success of a supervised framework relies on identifying suitable features for discriminating between causal and non-causal relations. The Bradford Hill causality considerations are a collection of nine factors that are often considered by experts to evaluate whether a drug and health outcome pair may correspond to an ADR [5–7]. Therefore, the Bradford Hill causality considerations seem an ideal basis for engineering suitable causal discriminative features to be used as input to train an ADR signalling classifier. The aim of this paper is to investigate whether such a classifier can be trained to successfully automate the

* Corresponding author.
E-mail address: jenna.reps@nottingham.ac.uk (J.M. Reps).

http://dx.doi.org/10.1016/j.jbi.2015.06.011
1532-0464/

process of using the Bradford Hill causality considerations to identify causality.

Our proposed supervised Bradford Hill's methodology is evaluated by considering the problem of signalling ADRs that occur shortly after being prescribed a medication. The data used in this study are from a large UK electronic healthcare database that contains medical records for millions of patients in the UK. The database is over 300 GB in size, therefore it is important to consider the efficiency of the feature engineering. The Bradford Hill's causality considerations were developed by an epidemiologist in the 1960s with experience in identifying causal relationships between drugs and health outcomes. They have been successfully implemented, by the process of manual review, as a means to determine causality in many epidemiological studies [8]. The considerations state that nine factors should be considered when assessing causality between a drug and health outcome. The factors are: (i) association strength, (ii) temporality, (iii) consistency, (iv) specificity, (v) biological gradient, (vi) experimentation, (vii) analogy, (viii) coherence and (ix) plausibility. As longitudinal observational databases contain data that can give insight into many of these considerations, we should take advantage of the data available to create a supervised signal detection framework that can imitate the causality review process.

The problem of identifying ADRs has often relied on the use of spontaneous reporting system (SRS) data. SRS data are composed of reported cases where somebody has suspected that a drug caused an ADR [9]. Common methods for detecting ADRs using SRS data are the disproportionality methods [10] that calculate a measure of association strength between the drug and health outcome based on inferring approximate background rates using all the reports. However, it is not possible to calculate the actual background incidence rates corresponding to the drug or health outcome using SRS data. Issues with under-reporting [11] can limit the ability to detect ADRs using SRS data and consequently, there has been an interest in using longitudinal observational data to aid ADR detection. Recent advances in using SRS data for signalling ADRs have focused on utilising all the SRS data and have considered non-association strength features [12,13]. It was shown that considering a variety of features lead to an improvement in ADR detection compared to standard methods [12]. However, this idea is currently unexplored for ADR detection using longitudinal observational databases, although there has been preliminary work suggesting Bradford Hill based features may add a new perspective for analysing electronic healthcare records [14].

Longitudinal observational data has been a recent focus of attention for extracting new drug safety knowledge due to it being a cheaper and often safer alternative to experimentation such as randomised controlled trials. Existing method for signalling ADRs using longitudinal observational databases include adapted disproportionality methods [15,16], association rule mining techniques [17,18], or adaptions of epidemiological studies [19]. All the large scale signalling methods are unsupervised, focus mostly on the measure of association strength and tend to have a high false positive rate in real life data [2,3], although some supervised techniques have been developed for specific cases. In [20], an ensemble technique combining simple epidemiology study designs to identify paediatric ADRs was shown to perform well. This suggested that incorporating supervised learning for ADR detection might lead to the improvement of signalling ADRs. For supervised learning to be fully utilised in this field, it is important to identify suitable features for the model. This motivates the idea of using a standard set of causal considerations widely implemented by experts in the field of epidemiology as a basis to engineer features. Numerous observational databases, including electronic healthcare records, tend to have hierarchies in the data recording [21,22]. It may be important to consider the hierarchies when searching for causal relationships because the relationship may be non-obvious when considering a high level item due to it occurring less frequently, but obvious when an abstract perspective is taken. If not taken into consideration, the hierarchal nature of the databases may weaken a signal. Therefore, we also propose features based on medical event coding hierarchies.

Outside of the field of drug safety, existing methods developed with the aim of identifying causal relationships within longitudinal observational data are often based on Bayesian networks [23]. Due to the complexity of creating a complete Bayesian network, many of the proposed methods are considered inappropriate for 'big' data [24]. However, constraint-based causal detection has been suggested as a means to handle 'big' data by applying metaheuristics that reduce the problem space [25]. Unfortunately these methods cannot overcome the common issues found within medical longitudinal data such as selection bias and do not consider hierarchal structures, and are therefore not currently suitable for signalling ADRs.

The continuation of this paper is as follows. Section 2 details the database used within this research and the proposed supervised Bradford Hill framework. In Section 3 we present the results of the supervised Bradford Hill framework's performance for signalling ADRs using a real database containing millions of UK patient records. The implications of the results are discussed in Section 4. The paper concludes with Section 5.

2. Materials and methods

2.1. THIN database

The data used in this paper were extracted from The Health Improvement Network (THIN) database, an electronic healthcare database containing UK primary care records for over 3.7 million active patients [26] (www.thin-uk.com). As the database contains time stamped records of medical events (e.g., myocardial infarction or vomiting) and drug prescriptions, each patient's medical state can be observed over time and temporal relationships between drugs and medical events can be identified. The THIN data used in this research contained over 200 million medical records and over 350 million prescription records.

The THIN database consists of heterogeneous data with multiple hierarchal structures. The database contains three key tables; the patient table, the medical table and the therapy table. For privacy reasons the patients' identities are not stored in the database, instead, each patient is assigned a unique reference known as the patientID that is used to determine which patient each record in the database corresponds to. The patient table contains information about each patient such as their date of birth, gender and date of registration or date of death (if they have died). The medical and therapy tables contain time stamped records of any medical or therapy event experienced by the patients, respectively. The database is normalised such that medical event descriptions and drug details are stored into separate tables and linked with unique references. The unique reference of a medical event is known as the Read code [22] and the unique reference of a drug is known as a drugcode.

The Read codes have a hierarchical coding system encompassing five levels of specificity, with level one Read codes representing very general events and level five Read codes representing very specific events. The level of a Read code is determined by its length. An example of a level one Read code is '1' and an example of a level 5 Read code is '11a1b'. The level 1 Read code 'G' is the parent of any Read code starting with 'G'. For example, the level 1 Read code 'G' representing the medical event 'Circulatory system disease', it is the parent of the Read codes:

Level 2 : 'G5' – 'Other forms of heart disease'.
Level 3 : 'G57' – 'Cardiac dysrhythmias'.