Computer Methods and Programs in Biomedicine 161 (2018) 25–38

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Supervised signal detection for adverse drug reactions in medication dispensing data

Tao Hoang[a,*], Jixue Liu[a], Elizabeth Roughead[b], Nicole Pratt[b], Jiuyong Li[a]

[a] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes Boulevard, South Australia 5095, Australia
[b] School of Pharmacy and Medical Sciences, University of South Australia, City East Campus, North Terrace Adelaide, South Australia 5001, Australia

ARTICLE INFO

Article history:
Received 24 October 2017
Revised 12 March 2018
Accepted 20 March 2018

Keywords:
Supervised machine learning
Gradient boosting
Drug
Adverse event
Signal detection
Adverse drug reaction
Medication dispensing data

ABSTRACT

Motivation: Adverse drug reactions (ADRs) are one of the leading causes of morbidity and mortality and thus should be detected early to reduce consequences on health outcomes. Medication dispensing data are comprehensive sources of information about medicine uses that can be utilized for the signal detection of ADRs. Sequence symmetry analysis (SSA) has been employed in previous studies to detect signals of ADRs from medication dispensing data, but it has a moderate sensitivity and tends to miss some ADR signals. With successful applications in various areas, supervised machine learning (SML) methods are promising in detecting ADR signals. Gold standards of known ADRs and non- ADRs from previous studies create opportunities to take into account additional domain knowledge to improve ADR signal detection with SML.
Objective: We assess the utility of SML as a signal detection tool for ADRs in medication dispensing data with the consideration of domain knowledge from DrugBank and MedDRA. We compare the best performing SML method with SSA.
Methods: We model the ADR signal detection problem as a supervised machine learning problem by linking medication dispensing data with domain knowledge bases. Suspected ADR signals are extracted from the Australian Pharmaceutical Benefit Scheme (PBS) medication dispensing data from 2013 to 2016. We construct predictive features for each signal candidate based on its occurrences in medication dispensing data as well as its pharmacological properties. Pharmaceutical knowledge bases including DrugBank and MedDRA are employed to provide pharmacological features for a signal candidate. Given a gold standard of known ADRs and non-ADRs, SML learns to differentiate between known ADRs and non-ADRs based on their combined predictive features from linked sources, and then predicts whether a new case is a potential ADR signal.
Results: We evaluate the performance of six widely used SML methods with two gold standards of known ADRs and non-ADRs from previous studies. On average, gradient boosting classifier achieves the sensitivity of 77%, specificity of 81%, positive predictive value of 76%, negative predictive value of 82%, area under precision-recall curve of 81%, and area under receiver operating characteristic curve of 82%, most of which are higher than in other SML methods. In particular, gradient boosting classifier has 21% higher sensitivity than and comparable specificity with SSA. Furthermore, gradient boosting classifier detects 10% more unknown potential ADR signals than SSA.
Conclusions: Our study demonstrates that gradient boosting classifier is a promising supervised signal detection tool for ADRs in medication dispensing data to complement SSA.

1. Introduction

Adverse drug reactions (ADRs) are unpleasant or harmful effects associated with taking a medicine [1]. For instance, anti-inflammation drugs such as ibuprofen have been known to be associated with gastrointestinal bleeding (stomach ulcers) [2]. ADRs are among the top five causes of hospitalizations and deaths in the U.S., costing billions of dollars annually [3,4]. Therefore, ADRs should be detected early to minimize consequences on health and cost. Clinical trials, however, are unable to identify all possible ADRs due to limited population sizes [5]. Thus, post- marketing drug safety surveillance, or pharmacovigilance, is necessary to continue the detection of ADRs in larger populations. Pharmacovigilance has mainly relied on spontaneous reporting systems (SRSs),

[*] Corresponding author.
E-mail address: tao.hoang@mymail.unisa.edu.au (T. Hoang).

https://doi.org/10.1016/j.cmpb.2018.03.021
0169-2607/© 2018 Elsevier B.V. All rights reserved.

Table 1
An example of the PBS medication dispensing dataset. (Note that this is for illustration only).

Patient ID	Transaction ID	Drug	ATC code	Supply date
1	1	tramadol	N02AX02	21/03/2014
1	2	aspirin	B01AC06	21/03/2014
1	3	candesartan	C09CA06	17/04/2014
2	6	aspirin	B01AC06	14/02/2012
2	7	candesartan	C09CA06	06/03/2012
2	8	metoclopramide	A03FA01	08/04/2012
3	10	lisinopril	C09AA03	09/01/2014
3	11	codeine	R05DA04	13/02/2014

e.g., U.S. Food and Drug Administration Adverse Event Reporting System. SRSs allow drug consumers, health professionals and pharmaceutical companies to report suspected ADRs. However, detecting ADRs from SRSs has several limitations. First, SRSs suffer from significant under-reporting. In fact, approximately 90% of ADRs may not have been reported to SRSs [6]. Second, the passive nature of SRSs may lead to reporting bias [5,7].

Given the limitations of SRSs, medication dispensing data have been utilized as a complementary source for ADR detection. Medication dispensing data contain all records of prescribed drugs with corresponding dates of supply generally collected for substantial populations, and thus are less likely to suffer from under-reporting or reporting bias [8]. An example of medication dispensing data is the Pharmaceutical Benefit Scheme (PBS) dataset in Australia [9]. Table 1 presents an example of the PBS dataset with three patients. While medication dispensing data do not usually include any record of health outcomes, drugs can be used as proxies for adverse events that they treat. For instance, the initiation of candesartan may indicate the treatment of emergent hypertension as an ADR whereas metoclopramide may indicate emergent nausea. Assuming that the time at which a drug is prescribed is close to the time the drug is taken by the patient, a patient's sequence of prescriptions approximately reflects the temporal order that the patient takes these drugs. Thus, medication dispensing data can be utilized to gain insights into the temporal relationships between prescriptions for detecting signals of ADRs. A prescription sequence $\langle \text{drug}_1 \rightarrow \text{drug}_2 \rangle$ may signal a potential adverse event indicated by drug_2 and potentially induced by drug_1 or it could be intended coadministration. For example, ⟨raloxifene→frusemide⟩ signals that raloxifene potentially induces oedema that is indicated by frusemide [10].

SSA has been used previously as an ADR signal detection tool for medication dispensing data. The principle behind SSA is to identify the asymmetry in the sequence of first prescriptions between two drugs within a time period [11,12]. The advantage of SSA is the consistent performance across different datasets [8,13]. While SSA has been shown to be robust, it is subject to several limitations. First, SSA was found to have a moderate sensitivity for detecting ADRs [10], i.e., tends to miss some ADR signals. In addition, to date, SSA has been used mainly for case-by-case assessment, i.e., the input signal candidate has been filtered and suspected to be potential ADR signal by medical experts [14]. The visual output of SSA allows domain experts to review the output and use their knowledge of drug indications, mechanism and onset of action and side effects to assess potential ADR signals [14]. The advent of extensive domain knowledge bases such as DrugBank and MedDRA provides the potential to automate the selection or filtering of potential ADR signals based on prespecified domain attributes.

With successful applications in various areas, supervised machine learning (SML) methods are promising in detecting ADR signals. In fact, SML has demonstrated effectiveness in many real-word applications of healthcare, marketing, spam detection, etc. [15]. Available gold standards of known ADRs and non-ADRs from previous studies [10,16] create opportunities to take into account additional domain knowledge such as drug indications from DrugBank and MedDRA to improve ADR detection with SML. Known ADRs are adverse events listed in the product information leaflets of particular drugs [16] or have been detected in clinical trials [10]. Non ADRs are those not listed in the product information of a drug and considered unlikely to be ADR signals by domain experts. Given a gold standard of known ADRs and non-ADRs, SML learns to differentiate between known ADRs and non-ADRs using their combined predictive features from linked sources, and then predicts whether a new case is a potential ADR signal.

In this study, we investigate the utility of SML and domain knowledge bases in detecting signals of ADRs from medication dispensing data. We model the ADR signal detection problem as a supervised machine learning problem by linking medication dispensing data with domain knowledge bases. We utilize the PBS dataset that contains medication dispensing records of all patients subsidized by the Australian government from 2013 to 2016 [9]. Our objective is to identify all sequences of the form $\langle \text{drug}_1 \rightarrow \text{drug}_2 \rangle$ that signal potential ADRs. We construct predictive features for each signal candidate based on the temporal relationships between drug_1 and drug_2 in the PBS dataset as well as their pharmacological properties. For instance, the number of patients with the first prescription of drug_2 K weeks after the first prescription of drug_1 was shown to be a useful temporal feature for ADR detection as its distribution over K tends to be similar for ADR signals [12]. Pharmacological features may also help improve the detection of ADRs by reducing spurious signals. For example, if drug_1 and drug_2 share many similar indications, this is likely to be coadministration or medication switching rather than a potential ADR signal. To utilize pharmacological features, we augment drugs in the PBS dataset with their pharmacological information from DrugBank [17] and MedDRA [18] via their anatomical therapeutic chemical (ATC) codes [19] and indication names.

We evaluate the performance of six commonly used SML methods using two gold standards containing known ADRs and non-ADRs [10,16] and one exploration set of known ADRs and unknown potential ADR signals [8]. While SML was employed in previous studies for ADR signal detection in the health improvement network (THIN) data [20,21], only random forests classifier [22] was studied and the data contain records of both prescriptions and health outcomes instead of just prescriptions as in our case. To the best of our knowledge, the performance of SML has not been studied on medication dispensing data. We found that gradient boosting classifier consistently outper- form SSA and other SML methods (i.e., logistic regression, decision tree, support vector machine, neural network, random forest) in most of the metrics across different gold standards. The average sensitivity, specificity, positive predictive values and negative predictive value, area under precision-recall curve, and area under receiver operating characteristic curve of gradient boosting classifier are 77%, 81%, 76%, 82%, 81%, and 82% respectively. Particularly, gradient boosting classifier has 21% higher sensitivity and comparable specificity with SSA. This suggests that gradient boosting classifier can be employed as an ADR signal detection tool to complement SSA and other existing methods.

2. Data

2.1. Medication dispensing dataset

We utilize the PBS medication dispensing dataset in Australia [9]. The data covers a random 10% sample of patients subsidized by the Australian government with their routinely updated records of prescribed drugs and corresponding dates of supply from 2013

Table 2
Statistics of our PBS medication dispensing dataset.

Statistics	Value
Total number of prescription transactions	7,294,244
Total number of patients	1,807,159
Total number of drugs	728
Total timespan of the data	4 years (Jan 2013–Dec 2016)

to 2016. An example of the dataset is shown in Table 1. Each patient has a set of transactions in ascending order of supply dates, i.e., from earliest to latest. Each transaction consists of a drug prescribed to a patient at a particular date of supply. Each drug is identified by a unique ATC code, e.g., C09CA06 for candesartan. For each drug prescribed to a patient, we only utilize its first prescription transaction and ignore the subsequent non-first transactions. To ensure all the drugs are newly prescribed, we also remove transactions with drugs dispensed between July 2012 and December 2012 as this was the start of the data and we were unable to distinguish new use from prevalent use in this period. Table 2 summarizes the statistics of our PBS medication dispensing dataset after the preprocessing. The dataset contains 7,294,244 prescription transactions from 1,807,159 patients, constituting a total of 728 unique drugs. In this study, we assume that the time at which a drug is prescribed to a patient is close to the time that drug is taken by the patient. As a result, a patient's sequence of prescribed drugs approximately reflects the temporal order that the patient takes these drugs.

2.2. Pharmacological knowledge bases

To utilize pharmacological features in the signal detection of ADRs, we need to access the information regarding drug indications, i.e., the medical conditions that can be treated by a particular drug. Each drug in the medication dispensing data was integrated with Structured Indications from DrugBank [17] via its ATC code. Furthermore, we enriched the drug indications by linking Structured Indications to hierarchies in the medical dictionary for regulatory activities (MedDRA) 19.0 [18]. In particular, each indication in structured indications was mapped to lowest level terms (LLTs), preferred terms (PTs), high level terms (HLTs), and high level group terms (HLGTs) in MedDRA. If there is an exact match between an indication's name (e.g., coughing) and a term in LLTs or PTs, the indication was firstly linked to the LLT term or PT term. Since each LLT term or PT term has corresponding terms in LLT, PTs, HLTs, and HLGTs, the indication is also linked to terms in all levels of MedDRA.

2.3. ADR gold standards

We evaluate our methods using two gold standards extracted from previous studies. The details of the gold standards are summarized in Table 3. The first gold standard, named Wahab13, consists of 67 known ADRs and 83 non-ADRs [10]. The details of Wahab13 can be found in the Appendix 2 and Appendix 3 of Wahab et al [10]. Known ADRs refer to those adverse events of particular medicines that were identified in randomized controlled trials. Non ADRs are those not listed in the product information leaflet of a medicine and any other medicine in the same therapeutic class and considered unlikely to be ADR signals by domain experts. Wahab13 can be used to both train and test SMLs. The second gold standard, Harpaz14, contains 58 known ADRs and 65 non-ADRs [16]. Similar to Wahab13, Harpaz14 is usable for both training and testing. The details of Harpaz14 can be found in the supplementary material of Harpaz et al. [16].

We choose Harpaz14 and Wahab13 for evaluation for three reasons. First, they are the two most recently published studies that include gold standards for known ADRs and non-ADRs. Second, the drugs in Wahab13 and Harpaz14 are diverse. Wahab13 covers drugs with high usage volume in Australia [10], while Harpaz14 contains diverse drugs of multiple types from the US Food and Drug Administration [16]. Lastly, the natures of known ADRs in Wahab13 and Harpaz14 are different. Known ADRs in Wahab13 are retrieved from randomized controlled trials, whereas known ADRs in Harpaz14 are listed in product labels of the US Food and Drug Administration. The differences between two datasets allow us to test the generalizability of our methods.

Besides the two gold standards, we also utilize an exploration set, Wahab16, to assess the ability of our methods in picking up unknown potential ADR signals. Wahab16 contains 41 known ADRs listed in the medicine product information leaflets and 65 unknown potential ADR signals [8]. Unknown potential ADRs are neither known ADRs nor non-ADRs. Unlike Wahab13 and Harpaz14, Wahab16 is used only for testing and exploration. All the known ADRs, non-ADRs and unknown potential ADR signals are encoded into the form $\langle \text{drug}_1 \rightarrow \text{drug}_2 \rangle$ by domain experts to be compatible with the medication dispensing dataset.

3. Methods

3.1. Overview

Fig. 1 presents the workflow of our approach. Since the medication dispensing dataset does not contain any record of adverse event or health outcome, drugs are used as proxies for adverse events that they treat. Given the medication dispensing dataset, our main goal is to identify a set of sequences of the form $\langle \text{drug}_1 \rightarrow \text{drug}_2 \rangle$ that signal potential ADRs. Particularly, the adverse event is indicated by drug_2 and potentially induced by drug_1. Our signal detection process consists of two main steps. First, we extracted all the sequences $\langle d_1 \rightarrow d_2 \rangle$ from the medication dispensing dataset such that for each of them, the drug d_2 occurs within the T_{ADR} time period after the drug d_1 in at least one patient. We set $T_{ADR} = 1$ year by default as it has been shown to be an appropriate time period for ADR signal detection in previous studies [8,10,23]. We showed empirically that one-year time period is the best option in the Results section. After the extraction of sequences, we computed the values of heterogeneous features for each sequence. The details of features will be discussed in a subsequent section. Given the sequences and their features, we employed a supervised ADR classifier (i.e., SML) to predict whether each sequence is a potential ADR signal. Lastly, we excluded signals that are known ADRs and retain unknown potential ADR signals for experts' further investigation.

The core of our approach is the supervised ADR classifier. We utilized the gold standards containing known ADRs and non-ADRs (i.e., Wahab13 and Harpaz14) to build the ADR classifier. Each gold standard was split into two parts. One part of the gold standard was used to train the classifier and the remaining part to test the performance of the trained classifier. The split was repeated multiple times to reduce the variance in the performance, which is referred to as cross validation [15]. Before the training and testing steps, the features of sequences were computed in a similar way as in the signal detection process. The following section describes the supervised ADR classifier in more detail, while the Results section presents its performance.

3.2. Supervised ADR classifiers

In this section, we describe the principles behind using SML to predict potential ADR signals. Suppose we have a training set of N

Table 3
Statistics of our gold standards.

Gold standard	Number of ⟨drug1 →drug2⟩ indicating known ADRs	Number of ⟨drug1 →drug2⟩ indicating non-ADRs
Wahab13 [10]	67	83
Harpaz14 [16]	58	65

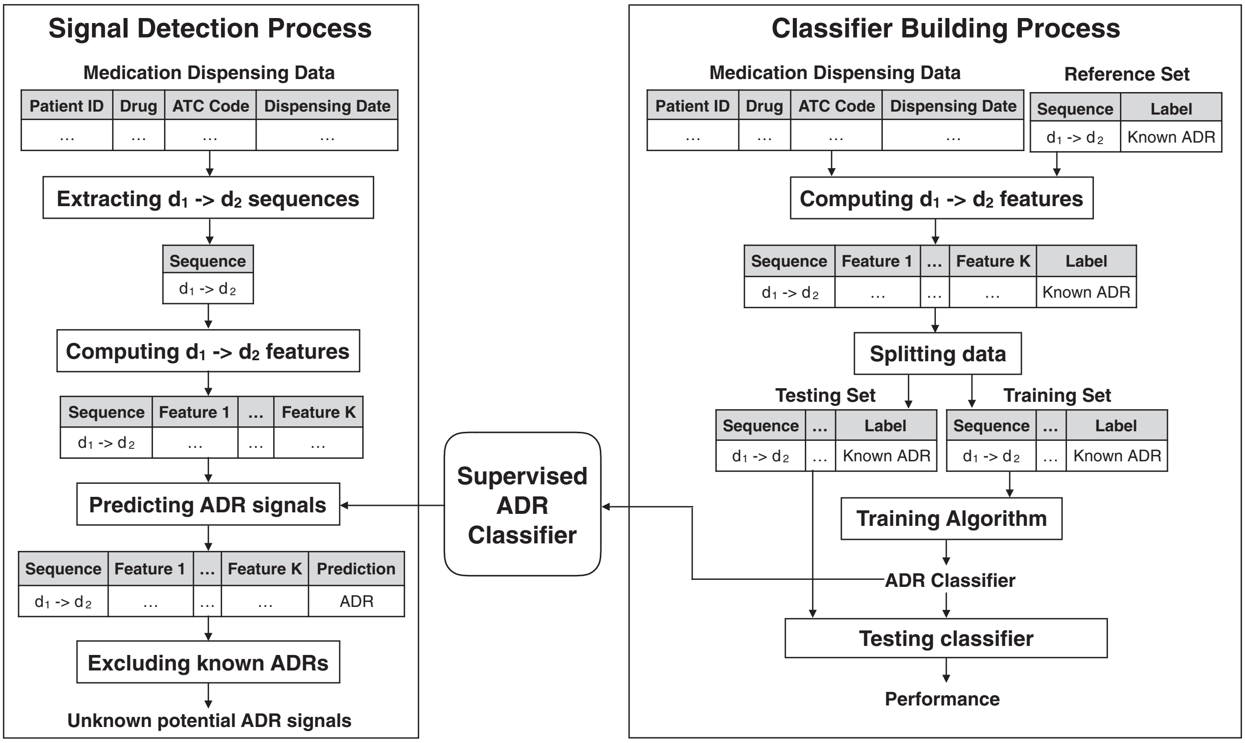

Signal Detection Process
Medication Dispensing Data
Patient ID | Drug | ATC Code | Dispensing Date
… | … | … | …
Extracting d₁ -> d₂ sequences
Sequence
d₁ -> d₂
Computing d₁ -> d₂ features
Sequence | Feature 1 | … | Feature K
d₁ -> d₂ | … | … | …
Predicting ADR signals
Sequence | Feature 1 | … | Feature K | Prediction
d₁ -> d₂ | … | … | … | ADR
Excluding known ADRs
Unknown potential ADR signals
Supervised ADR Classifier
Classifier Building Process
Medication Dispensing Data
Patient ID | Drug | ATC Code | Dispensing Date
… | … | … | …
Reference Set
Sequence | Label
d₁ -> d₂ | Known ADR
Computing d₁ -> d₂ features
Sequence | Feature 1 | … | Feature K | Label
d₁ -> d₂ | … | … | … | Known ADR
Splitting data
Testing Set
Training Set
Sequence | … | Label
d₁ -> d₂ | … | Known ADR
Sequence | … | Label
d₁ -> d₂ | … | Known ADR
Training Algorithm
ADR Classifier
Testing classifier
Performance

Fig. 1. The workflow of our approach.

observations $\{(x_1,y_1),\ldots,(x_N,y_N)\}$ where x_i is the feature vector of the i-th sequence $\langle d_1^{(i)} \rightarrow d_2^{(i)} \rangle$ and $y_i \in \{0, 1\}$ indicates whether $\langle d_1^{(i)} \rightarrow d_2^{(i)} \rangle$ is a known ADR or non-ADR. The goal of classifier learning is to find the best estimate of the prediction function $f(x)$ mapping an input feature vector x to the binary output y by minimizing the expectation of some loss function $L(y, f(x))$ over the join distribution of $\{x_i,y_i\}^N_{i=1}$ [15,24] (training phase).

$$\hat{f}(x) = \underset{f(x)}{\operatorname{argmin}}\ E_{y,x}\ L(y, f(x)) \qquad (1)$$

The prediction function $f(x)$ combines values of features in a particular way. Different classifiers have different forms of $f(x)$. The estimated function $\hat{f}(x)$ can be used to predict y on observations where only the feature vectors x are available (testing phase).

In this study, we compared the performance of six widely used classifiers in detecting signals of ADRs: logistic regression [25], decision tree [26], support vector machine [27], neural network [28], random forests [22], and gradient boosting [29,30]. We employed the Scikit-learn library [31] in Python to provide the implementations for all the classifiers. In the following, we briefly describe each classifier and its best configuration in our study.

- Logistic regression: logistic regression classifier is one of the most widely used classification methods. Logistic regression classifier utilizes logistic function as the prediction function $f(x)$ to linearly combine features for prediction.
- Decision tree: decision tree classifier uses the tree structure to represent the prediction function $f(x)$ for decision making. Each internal node in the tree indicates a test on the value of a feature while each outgoing branch from the node indicates the outcome of the test. Each leaf node tells whether the preceding path of feature tests indicates a potential ADR signal or not. The idea of constructing a decision tree is to iteratively select an unused feature whose values best split the sub-dataset in a path into two classes based on certain criterion. In this study, we use the Gini impurity criterion [32] to select the feature at each node.
- Support vector machine: support vector machine classifier represents known ADRs and non-ADRs as points in space according to their feature values. The idea of support vector machine is to identify $f(x)$ that well separates known ADRs from non-ADRs in space. Support vector machine can be categorized into linear support vector machine and non-linear support vector machine. In this study, we utilize linear support vector machine.
- Neural network: neural network classifier is a classification method whose structure $f(x)$ contains layers of neurons. There are three types of layers: input layer, hidden layer, and output layer. There is one input layer containing all the features for ADR classification. There is one output layer that holds the value of prediction function $f(x)$. Between input and output layer are zero or multiple hidden layers. There are one or multiple neurons in each hidden layer. Features and the output are special types of neurons. For our study, we only consider feed-forward neural network [33], in which a neuron in each layer is connected to neurons in the next layer. Information from the input layer is propagated through the hidden layers to the output layer. Each neuron is the output of an activation function taking the weighted linear summation of neuron values in the previous layer as input. Various activation functions including

identity function, logistic function, etc. have been utilized. Here we use 1 hidden layer with 100 neurons and rectified linear unit function as the activation function.

- Random forests: random forests classifier is an ensemble method that combines multiple ADR decision tree classifiers. The intuition is to obtain a strong ADR classifier out of many weak ADR classifiers. Each ADR decision tree classifier in random forests is trained using a different subset of training data that are randomly sampled with replacement. Random forests classifier then outputs the prediction that has the most votes based on all the ADR decision tree classifiers. We empirically build the ADR random forests classifier with 100 ADR decision tree classifiers. Each ADR decision tree classifier uses the Gini impurity criterion to select the feature at each node.
- Gradient boosting: gradient boosting classifier is another ensemble classifier with similar intuition as random forests classifier. Gradient boosting aggregates many weak ADR classifiers into a strong ADR classifier. Gradient boosting begins with learning a base weak ADR classifier and then iteratively learning additional weak classifiers that focus on observations difficult to predict in previous iterations. Different from random forests classifier, a weak ADR classifier in gradient boosting classifier can be any supervised machine learning classifier such as logistic regression, decision tree, etc. [15], which alone does not achieve high predictive per- formance. Another fundamental difference between gradient boosting classifier and random forests classifier is that gradient boosting classifier focuses on specific difficult-to-predict observations after each iteration instead of randomly sampled observations as in random forests classifiers. In this study, we empirically utilize 1000 decision tree classifiers as weak ADR classifiers. Since gradient boosting classifier achieves the best ADR prediction performance as we shall see in the Results section, we describe more details about gradient boosting classifier in the Appendix.

3.3. Sequence features

We now describe the features based on which each sequence $\langle d_1 \rightarrow d_2 \rangle$ is classified as whether it signals an ADR. The features are categorized into three groups: (1) statistic features, (2) group-based statistic features, and (3) pharmacological features.

3.3.1. Statistic features

This group of features captures the prevalence of different configurations of the sequence in the medication dis- pensing dataset.

- Sequence support: the number of patients to whom the first prescription of d_2 occurs within $T_{ADR} = 1$ year after the first prescription of d_1.
- Reverse sequence support: the number of patients to whom the first prescription of d_1 occurs within $T_{ADR} = 1$ year after the first prescription of d_2.
- d$_1$'s support: the number of patients to whom d_1 was prescribed.
- d$_2$'s support: the number of patients to whom d_2 was prescribed.
- 1st-week sequence support: the number of patients to whom the first prescription of d_2 occurs in the 1st week after the first prescription of d_1.
- 2nd-week sequence support: the number of patients to whom the first prescription of d_2 occurs in the 2nd week after the first prescription of d_1.
- …
- 52nd-week sequence support: the number of patients to whom the first prescription of d_2 occurs in the 52nd week after the first prescription of d_1.
- 1st-week reverse sequence support: the number of patients to whom the first prescription of d_1 occurs in the 1st week after the first prescription of d_2.
- 2nd-week reverse sequence support: the number of patients to whom the first prescription of d_1 occurs in the 2nd week after the first prescription of d_2.
- …
- 52nd-week reverse sequence support: the number of patients to whom the first prescription of d_1 occurs in the 52nd week after the first prescription of d_2.

Note that the features were computed up to the 52nd week as one year corresponds to 52 weeks. When T_{ADR} changes, the number of features also changes accordingly. For instance, if $T_{ADR} = 9$ months then the features are up to 39 weeks.

3.3.2. Group-based statistic features

Some drugs are so rarely prescribed in the medication dispensing dataset that it is difficult to pick up ADR signals associated with them. For instance, betaxolol was prescribed to only 243 patients and the sequence support of ⟨betaxolol→frusemide⟩ is only 7. While ⟨betaxolol→frusemide⟩ signals a known ADR ⟨betaxolol→oedema⟩ [8], it is not detected by SSA due to the lack of data. To alleviate the data rarity, we utilized additional features related to groups of drugs. Drugs having the same ATC fourth level (i.e., first five letters) belong to the same chemical subgroup and thus share many essential properties [19]. Let D_1 and D_2 represent the groups of drugs having the same ATC fourth levels with d_1 and d_2 respectively. Statistic features of $\langle D_1 \rightarrow D_2 \rangle$ might be helpful for detecting $\langle d_1 \rightarrow d_2 \rangle$. The group of drugs having the same ATC fourth level with betaxolol has the support of 11,523 patients.

- Group-based sequence support: the number of patients to whom the first prescription of D_2 occurs within $T_{ADR} = 1$ year after the first prescription of D_1.
- Group-based reverse sequence support: the number of patients to whom the first prescription of D_1 occurs within $T_{ADR} = 1$ year after the first prescription of D_2.
- D$_1$'s support: the number of patients to whom D_1 was prescribed.
- D$_2$'s support: the number of patients to whom D_2 was prescribed.
- Group-based 1st-week sequence support: the number of patients to whom the first prescription of D_2 occurs in the 1st week after the first prescription of D_1.
- Group-based 2nd-week sequence support: the number of patients to whom the first prescription of D_2 occurs in the 2nd week after the first prescription of D_1.
- …
- Group-based 52nd-week sequence support: the number of patients to whom the first prescription of D_2 occurs in the 52nd week after the first prescription of D_1.
- Group-based 1st-week reverse sequence support: the number of patients to whom the first prescription of D_1
- occurs in the 1st week after the first prescription of D_2.
- Group-based 2nd-week reverse sequence support: the number of patients to whom the first prescription of
- D_1 occurs in the 2nd week after the first prescription of D_2.
- …
- Group-based 52nd-week reverse sequence support: the number of patients to whom the first prescription of
- D_1 occurs in the 52nd week after the first prescription of D_2.

3.3.3. Pharmacological features

This group of features helps identify sequences that are likely to represent coadministration or medication switching rather than potential ADR signals. If $\langle d_1 \rightarrow d_2 \rangle$ is coadministration or medication switching, d_1 and d_2 tend to share similar indications or overlaps in prefixes of ATC codes. We employed DrugBank [17] to provide the indica- tions for each drug in the medication dispensing data. We also utilized hierarchy-based indications in MedDRA [18] as discussed in the Data section.

- Overlapping ATC prefixes: the number of prefixes shared by d_1 and d_2 in their ATC codes.
- Overlapping DrugBank indications: the number of DrugBank indications shared by d_1 and d_2.
- Overlapping MedDRA LLT indications: the number of MedDRA lowest level terms shared by d_1 and d_2.
- Overlapping MedDRA PT indications: the number of MedDRA preferred terms shared by d_1 and d_2.
- Overlapping MedDRA HLT indications: the number of MedDRA high level terms shared by d_1 and d_2.
- Overlapping MedDRA HLGT indications: the number of MedDRA high level group terms shared by d_1 and d_2.

3.4. Evaluation measures for ADR classifiers

We evaluated the performance of supervised ADR classifiers using sensitivity, specificity, positive predictive value, and negative predictive value [10]. Sensitivity measures the proportion of known ADRs in the gold standard that are predicted as potential ADR signals by the classifier. Specificity is the proportion of non-ADRs in the gold standard that are not predicted as potential ADR signals by the classifier. Positive predictive value measures the proportion of ADR signals predicted as potential ADR signals by the classifier that are actually known ADRs in the gold standard. Negative predictive value is the proportion of ADR signals not predicted as potential ADR signals by the classifier that are actually non-ADRs in the gold standard. All four measures are important to evaluate the methods. While sensitivity and specificity measure how well the methods correctly detect known ADRs and non-ADRs in the gold standard, positive predictive value and negative predictive value measure the ability of the methods in predicting whether a pair is a potential unknown ADR or spurious. These four measures have been used in previous studies on ADR signal detection such as Wahab et al [10].

Furthermore, we utilized the receiver operating characteristic curve (ROC curve) and the precision-recall curve (PR curve) to compare different classifiers. Each SML classifier predicts an ADR with a probability, and the most intuitive and widely used probability threshold to determine whether a sequence is a potential ADR signal is 0.5. ROC curve is generated by plotting the sensitivity against 1-specificity at different probability thresholds of the classifier (e.g., 0, 0.01, 0.02, . . . , 1). PR curve is created by plotting the precision (i.e., positive predictive value) against the recall (i.e., sensitivity) at different thresholds. The area under ROC curve (ROC–AUC) indicates the balance between sensitivity and specificity, while the area under PR curve (PR–AUC) the balance between precision and recall. These two curves have been intensively utilized for evaluating classification models [34].

3.5. SSA–baseline ADR signal detection tool

We compared the performance of SML methods with SSA, a current ADR signal detection tool in medication dispensing data [10–12,35–55]. Petri et al. was the first to introduce SSA in 1988 [11]. Hallas then conceptualized SSA to test the association between cardiovascular medications and depression in 1996 [12]. The principle behind SSA is to identify the asymmetry in the sequence of first prescriptions between two drugs d_1 and d_2 within a time period T_{ADR}. If d_1 induces the prescription of d_2 as a result of an ADR, patients with the first prescription of d_1 before the first prescription of d_2 are expected to outnumber patients for whom d_2 are firstly prescribed before d_1 are firstly prescribed. As a result, the crude sequence ratio was defined as the ratio between the number of patients with $\langle d_1 \rightarrow d_2 \rangle$ and the number of patients with $\langle d_2 \rightarrow d_1 \rangle$ [12]. The crude sequence ratio could be utilized as an estimate of the incident rate ratio of the ADR when d_1 is exposed versus when d_1 is not exposed [12,14]. While the crude sequence ratio is not affected by confounders that are stable over time, it is sensitive to changes in prescription trends [13]. For instance, if the use of d_2 increases due to changes in reimbursement, the number of patients with $\langle d_1 \rightarrow d_2 \rangle$ rises. In this case, the crude sequence ratio overestimates the true incident rate ratio and may be biased [14]. To eliminate this bias, Hallas proposed to divide the crude sequence ratio by the null-effect sequence ratio that captures the prescription trends in the background population [12]. The null-effect sequence ratio is computed as the expected sequence ratio of d_2 being firstly prescribed after d_1 is firstly prescribed if d_1 and d_2 are independent. Tsiropoulos et al. later modified the null-effect sequence ratio to restrict the time period between the first prescriptions of d_1 and d_2 [35]. Thus, the adjusted sequence ratio was defined as the ratio between crude sequence ratio and null-effect sequence ratio. SSA relies on the adjusted sequence ratio to predict potential ADR signals. If the 95% confidence interval lower limit of the adjusted sequence ratio exceeds one, $\langle d_1 \rightarrow d_2 \rangle$ is considered a potential ADR signal. The advantage of SSA is its consistent performance across different datasets [8,13].

4. Results

4.1. Validation against known ADRs and non-ADRs

In this section, we compare the performances of six supervised ADR classifiers upon detecting known ADRs and non-ADRs in the gold standards Wahab13 and Harpaz14. First, we describe the validation setting. Then we study the effect of varying the probability thresholds, time periods T_{ADR} and the set of features.

4.1.1. Validation settings

We designed the following four settings to validate the performances of supervised ADR classifiers in this study:

- Wahab13 (Cross Validation): the gold standard Wahab13 is split into a training set (75%) and a testing set (25%). The performances of each ADR classifier are averaged over 100 random splits.
- Harpaz14 (Cross Validation): the gold standard Harpaz14 is split into a training set (75%) and a testing set (25%). The performances of each ADR classifier are averaged over 100 random splits.
- Wahab13 (Train) + Harpaz14 (Test): each ADR classifier is trained using Wahab13 and tested with Harpaz14.
- Harpaz14 (Train) + Wahab13 (Test): each ADR classifier is trained using Harpaz14 and tested with Wahab13.

4.1.2. The performances of supervised ADR classifiers in validation

Table 4 presents the performances of six supervised ADR classifiers and SSA in terms of sensitivity, specificity, positive predictive value and negative predictive value under various validation settings. Figs. 2 and 3 show the PR curves and ROC curves with corresponding AUCs for different supervised ADR classifiers. Overall, gradient boosting classifier achieves the sensitivity of 77%, specificity of 81%, positive predictive value of 76%, negative predictive value of 82%, PR-AUC of 81%, and ROC-AUC of 82%, most of which are highest among other supervised ADR classifiers and SSA across different settings. In comparison with SSA, gradient boosting classifier improves the sensitivity by 21%, specificity by 3%, positive

Table 4
Performances of supervised ADR classifiers and SSA in validation.

Setting	Method	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Wahab13 (Cross validation)	Logistic regression	66%	78%	72%	74%
	Decision tree	71%	79%	74%	78%
	Support vector machine	56%	60%	59%	59%
	Neural network	64%	50%	52%	62%
	Random forests	66%	90%	85%	77%
	Gradient boosting	84%	87%	85%	87%
Harpaz14 (Cross validation)	Logistic regression	73%	72%	74%	74%
	Decision tree	74%	70%	72%	74%
	Support vector machine	53%	49%	51%	51%
	Neural network	49%	48%	44%	49%
	Random forests	68%	82%	80%	72%
	Gradient boosting	77%	80%	81%	79%
Wahab13 (Train) + Harpaz14 (Test)	Logistic regression	51%	63%	59%	55%
	Decision tree	74%	54%	63%	67%
	Support vector machine	67%	41%	55%	55%
	Neural network	70%	24%	49%	43%
	Random forests	44%	80%	70%	58%
	Gradient boosting	76%	75%	65%	83%
Harpaz14 (Train) + Wahab13 (Test)	Logistic regression	59%	73%	64%	69%
	Decision tree	48%	63%	51%	59%
	Support vector machine	67%	20%	41%	43%
	Neural network	52%	69%	58%	64%
	Random forests	49%	89%	79%	68%
	Gradient boosting	70%	80%	72%	79%
Wahab13 (Test)	SSA	64%	75%	68%	72%
Harpaz14 (Test)		47%	80%	61%	71%
Average	Logistic regression	62%	72%	67%	68%
	Decision tree	67%	67%	65%	70%
	Support vector machine	61%	43%	52%	52%
	Neural network	59%	48%	51%	55%
	Random forests	57%	85%	79%	69%
	Gradient boosting	77%	81%	76%	82%
	SSA	56%	78%	65%	72%

predictive value by 11%, and negative predictive value by 10% on average. This shows that gradient boosting classifier is able to detect additional potential ADR signals that might be unobserved by SSA without picking up more spurious signals. While the same gold standard Wahab13 is used in Wahab et al. [10] and in this paper, the performance of SSA is different for two reasons. First, the periods of medication dispensing data are different. Wahab et al. [10] used the records between 2000 and 2010, whereas we used the records between 2013 and 2016. Second, Wahab et al. [10] utilized hospitalization records in addition to medication dispensing records as in our case.

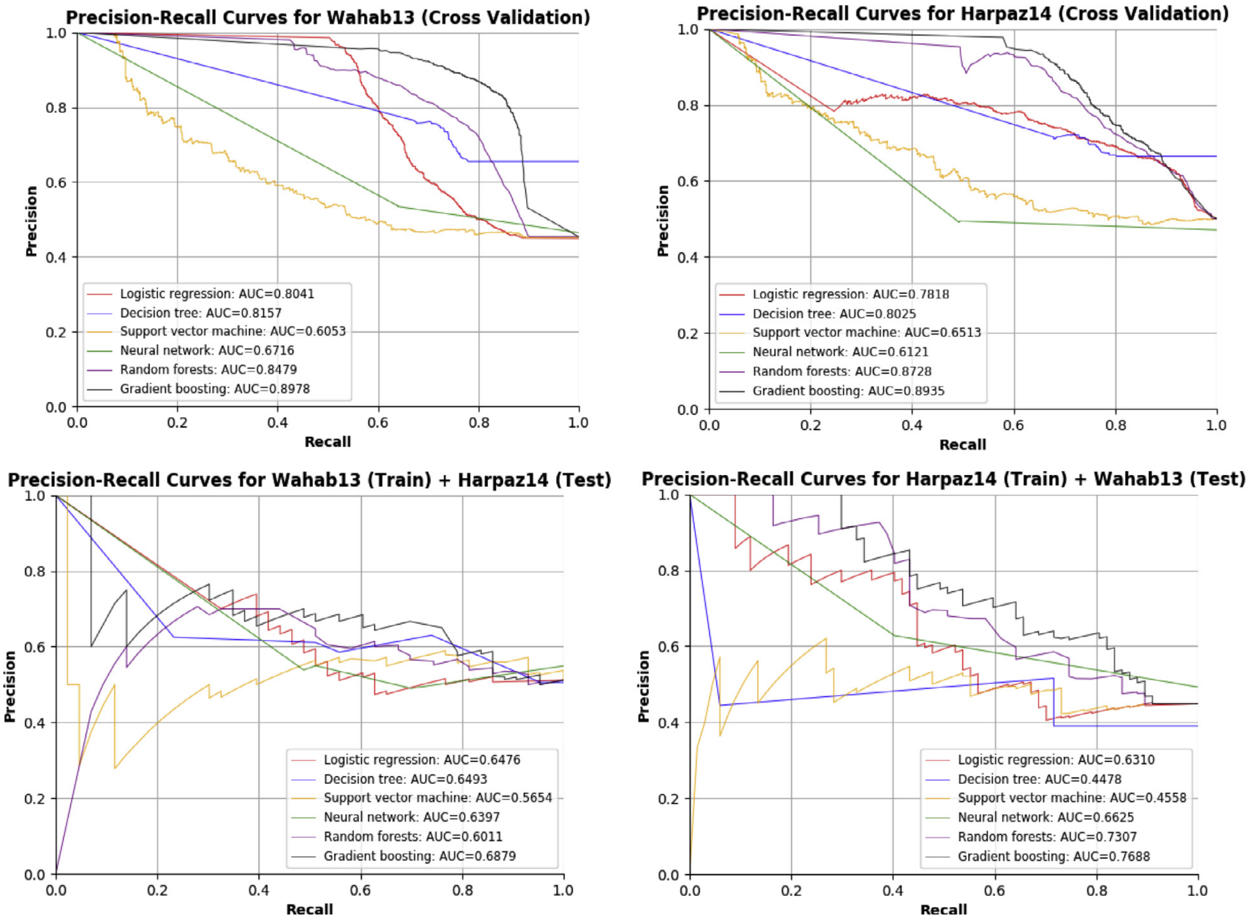
In addition, we observe that gradient boosting classifier consistently outperforms all other supervised ADR classi- fiers except random forests classifier in all six measures. Compared to random forests classifier, on average, gradient boosting classifier has 20% higher sensitivity, 13% higher negative predictive value, 5% higher PR-AUC, and 4% higher ROC–AUC, but 4% lower specificity and 3% lower positive predictive value. The slight trade-offs in specificity and positive predictive value allows gradient boosting classifier to detect more potential ADR signals than RF classi- fier, which result in much higher sensitivity and negative predictive value. Figs. 2 and 3 also demonstrated that gradient boosting classifier outperforms random forests classifier and other supervised ADR classifiers under dif- ferent probability thresholds. These results show that gradient boosting classifier is a promising ADR signal detection tool in medication dispensing data. Furthermore, in real-world applications, a supervised ADR classifier would al- ways be trained on one gold standard but used to detect unknown ADR signals, which are completely different from those on which it is trained. Thus, the encouraging performance of gradient boosting classifier when training on one gold standard and testing on another reflects its likely real-world applicability.

4.1.3. The effect of different probability thresholds

Table 5 presents the performance of gradient boosting classifier under different probability thresholds from 0.5 to 0.9. As the threshold increases, the sensitivity decreases and the specificity increases consistently. This is because fewer sequences are classified as potential ADR signals with higher thresholds. Thus, the number of true positives either remains unchanged or decreases, making the sensitivity unchanged or decrease. Likewise, as the number of false positives is either the same or reduced, the number of true negatives stays still or rises, and so is the specificity. In addition, we observe that the positive predictive value rises while the negative predictive value drops as the threshold increases. This demonstrates that higher thresholds can eliminate many false positives but at the same time introduce many false negatives. Furthermore, the results show that gradient boosting classifier still outperforms SSA on average under higher thresholds. For instance, when the threshold is 0.9, the gradient boosting classifier achieves the sensitivity of 64%, specificity of 87%, positive predictive value of 79%, and negative predictive value of 76%, which is higher than SSA with sensitivity of 56%, specificity of 78%, positive predictive value of 65%, and negative predictive value of 72%.

4.1.4. *The effect of time period T_{ADR}*

Table 6 demonstrates how the performance of gradient boosting classifier changes when the time period T_{ADR} varies from 12 months to 3 months. The results show that gradient boosting classifier obtains the best balance between sensitivity, specificity, positive predictive value, and negative predictive value when $T_{ADR} = 12$ months across different settings. On average, gradient boosting classifier also achieves the best performance when $T_{ADR} = 12$ months. In addition, as the time period T_{ADR} decreases, the sensitivity of the method also drops in each setting. This is because fewer sequences are extracted from the medication dis-

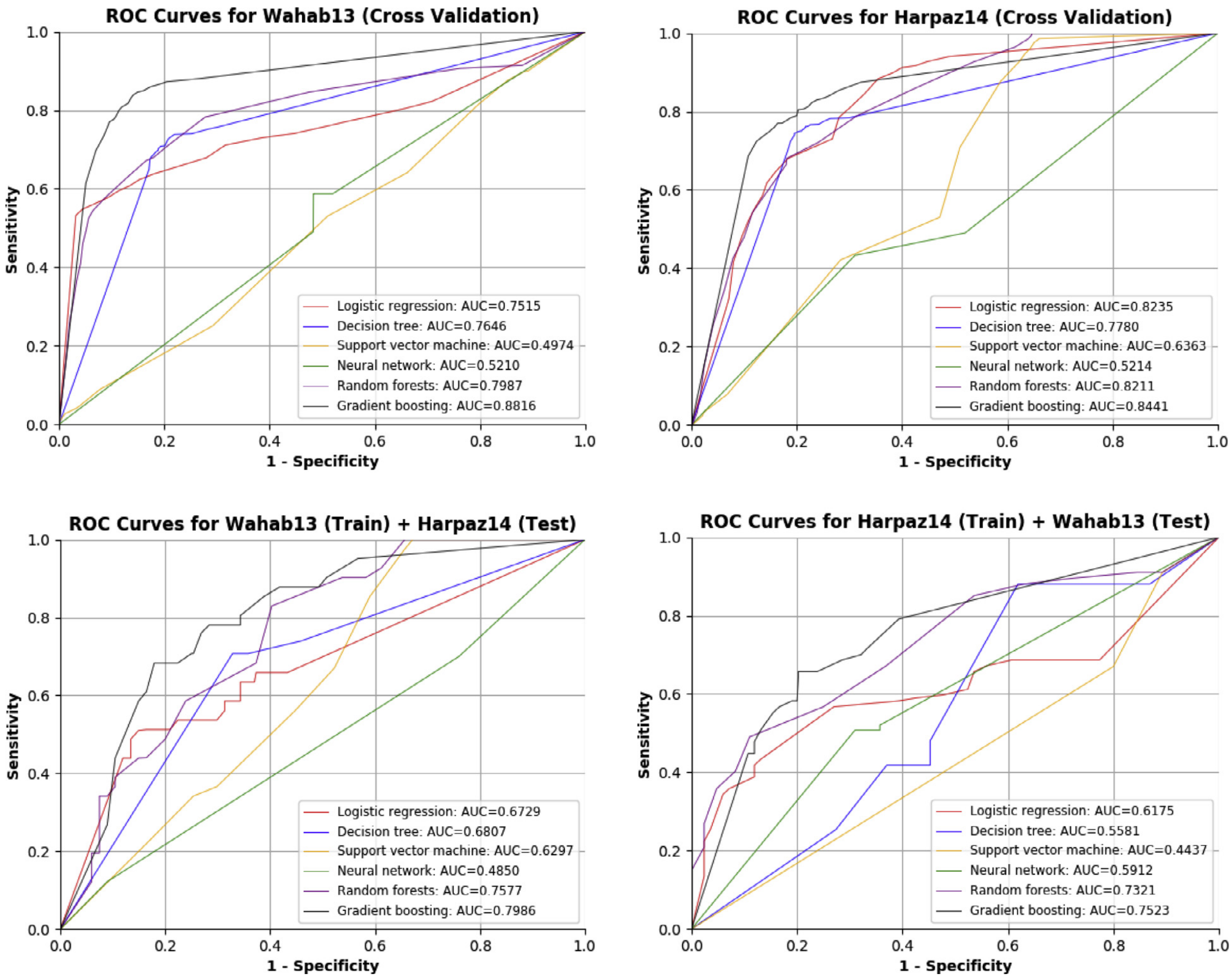


	Logistic regression	Decision tree	Support vector machine	Neural network	Random forests	Gradient boosting
Average PR-AUC	72%	68%	57%	65%	76%	81%

Fig. 2. Precision-recall curves for all validation settings.

Table 5
Performance of gradient boosting classifier under different probability thresholds.

Setting	Probability threshold	Negative	Specificity	Positive predictive value	Negative predictive value
Wahab13 (Cross validation)	0.5	84%	87%	85%	87%
	0.6	82%	88%	86%	86%
	0.7	81%	89%	86%	86%
	0.8	80%	91%	88%	86%
	0.9	78%	91%	89%	84%
Harpaz14 (Cross validation)	0.5	77%	80%	81%	79%
	0.6	76%	81%	81%	79%
	0.7	75%	81%	82%	77%
	0.8	74%	82%	82%	77%
	0.9	72%	86%	85%	76%
Wahab13 (Train) + Harpaz14 (Test)	0.5	76%	75%	65%	83%
	0.6	75%	76%	66%	82%
	0.7	71%	76%	66%	81%
	0.8	71%	78%	66%	81%
	0.9	61%	82%	68%	77%
Harpaz14 (Train) + Wahab13 (Test)	0.5	70%	80%	72%	79%
	0.6	63%	80%	72%	73%
	0.7	55%	82%	72%	70%
	0.8	51%	83%	73%	68%
	0.9	45%	88%	75%	67%
Average	0.5	77%	81%	76%	82%
	0.6	74%	81%	76%	80%
	0.7	71%	82%	77%	79%
	0.8	69%	84%	77%	78%
	0.9	64%	87%	79%	76%



	Logistic regression	Decision tree	Support vector machine	Neural network	Random forests	Gradient boosting
Average PR-AUC	72%	70%	55%	53%	78%	82%

Fig. 3. Receiver operating characteristic curves for all validation settings.

Table 6
Performance of gradient boosting classifier across different time period *T ADR*.

Setting	TADR (months)	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Wahab13 (Cross validation)	12	84%	87%	85%	87%
	9	83%	87%	85%	86%
	6	82%	88%	85%	86%
	3	82%	88%	86%	86%
Harpaz14 (Cross validation)	12	77%	80%	81%	79%
	9	20%	80%	47%	49%
	6	19%	81%	47%	49%
	3	17%	83%	48%	50%
Wahab13 (Train) + Harpaz14 (Test)	12	76%	75%	65%	83%
	9	40%	66%	55%	51%
	6	35%	66%	52%	49%
	3	33%	71%	54%	50%
Harpaz14 (Train) + Wahab13 (Test)	12	70%	80%	72%	79%
	9	65%	77%	70%	73%
	6	61%	64%	58%	67%
	3	61%	65%	59%	67%
Average	12	77%	81%	76%	82%
	9	52%	78%	64%	65%
	6	49%	75%	61%	63%
	3	48%	77%	62%	63%

Table 7
Performance of gradient boosting classifier for different features.

Setting	Features	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Wahab13 (Cross validation)	All features	84%	87%	85%	87%
	No statistic features	82%	87%	84%	86%
	No group-based statistic features	74%	84%	80%	80%
	No pharmacological features	82%	86%	84%	86%
Harpaz14 (Cross validation)	All features	77%	80%	81%	79%
	No statistic features	75%	75%	76%	76%
	No group-based statistic features	69%	58%	63%	65%
	No pharmacological features	75%	73%	75%	75%
Wahab13 (Train) + Harpaz14 (Test)	All features	76%	75%	65%	83%
	No statistic features	74%	44%	58%	62%
	No group-based statistic features	49%	76%	68%	58%
	No pharmacological features	70%	44%	57%	58%
Harpaz14 (Train) + Wahab13 (Test)	All features	70%	80%	72%	79%
	No statistic features	69%	72%	67%	74%
	No group-based statistic features	75%	48%	54%	71%
	No pharmacological features	56%	68%	59%	65%
Average	All features	77%	81%	76%	82%
	No statistic features	75%	70%	71%	75%
	No group-based statistic features	67%	67%	66%	69%
	No pharmacological features	71%	68%	69%	71%

pensing data. Furthermore, it can be observed that the gradient boosting classifier is generally more sensitive to time period T_{ADR} in Harpaz14 rather than in Wahab13. When T_{ADR} decreases from 12 months to 9 months, the performance of gradient boosting classifier drops significantly from 77% to 20% in Harpaz14 (Cross validation) while only 84% to 83% in Wahab13 (Cross validation). This suggests that Wahab13 is a better training set than Harpaz14, which may be because the known ADRs in Wahab13 comes from randomized controlled trials instead of product information leaflets as in Harpaz14.

4.1.5. The effect of different features

First, we examined the effect of different feature groups, i.e., statistic features, group-based statistic features, and pharmacological features. Table 7 shows the changes in the performance of gradient boosting classifier as different groups of features are missing in various settings. On average, across different settings, the performance of gradient boosting classifier drops when any type of feature is excluded. When statistic features are missing, the sensitivity, specificity, positive predictive value and negative predictive drop by 2%, 11%, 5%, and 7% respectively. When there is no group-based statistic features, the sensitivity, specificity, positive predictive value and negative predictive are reduced by 10%, 14%, 10%, and 13%. When pharmacological features are absent, the sensitivity, specificity, positive predictive value and negative predictive decrease by 6%, 13%, 7%, and 11%. These results show that all three types of features are essential in distinguishing between known ADRs and non-ADRs. Among three types of features, the performance of gradient boosting classifier changes most significantly when group-based statistic features are excluded. In other words, group-based statistic features are most important in the signal detection of ADRs. This may be because that drugs within the same group (i.e., same ATC fourth level) often have common ADRs and thus group-based statistic features are essential in signaling ADRs in which drugs are rarely prescribed.

Furthermore, we examined the effects of different individual features. Fig. 4 presents the top 50 features with highest relative feature importance in predicting potential ADR signals by gradient boosting classifier. The features are sorted by their relative importance descendingly. How to compute feature importance are discussed in [56] and implemented by the Scikit-learn library [31]. In brief, for each ADR decision tree classifier in gradient boosting classifier, we computed the importance of a feature as the proportion of observations it can be used to differentiate between known ADRs and non-ADRs. Then the importance of each feature is averaged over all ADR decision tree classifiers. Feature importance is in the range [0,1] with higher scores indicating more important features. Features belonging to different groups are colored differently. It can be observed that most features in the top 50 are group-based statistic features, i.e., orange bars. This is consistent with our earlier results in Table 7, meaning that group-based statistic features play the most important role in differentiating between known ADRs and non-ADRs. In particular, D_2's support and D_1's support are the two most important features. Intuitively, large values of D_2's support and D_1's support show that D_1 and D_2 are commonly prescribed to patients, and thus the signal $\langle d_1 \rightarrow d_2 \rangle$ is likely to be spurious. The same intuition applies for d_1's support and d_2's support and therefore they are also in the top 5. In addition, we observe that the group-based 1st-week sequence support and group-based 1st-week reverse sequence support are very important features (i.e., in the top 5). This most likely reflects the acute ADR response soon after initiating treatment. Two pharmacological features in the top 50 are overlapping ATC prefixes (rank 8th) and overlapping MedDRA HLT indications (rank 50th). Since drugs sharing similar indications often have a first few overlapping ATC letters, the overlapping ATC prefixes feature is important in distinguishing known ADRs from coadministrations or drug switching. The importance of the overlapping MedDRA HLT indications feature shows that high level terms in MedDRA are most appropriate to represent drug indications in ADR signal prediction.

4.2. Comparison of ADR signals detected by gradient boosting classifier and SSA

In this section, we compared the ADR signals detected by gradient boosting classifier and SSA from the testing set Wahab16. Among 106 signals in Wahab16, 41 are known ADRs and 65 are unknown potential ADRs. Fig. 5 depicts known ADRs and unknown potential ADR signals with their adjusted sequence ratios (by SSA) and probabilities (by gradient boosting classifier). Blue circles represent known ADRs while red squares indicate unknown potential ADR signals. A signal is picked up by SSA if the 95% confidence interval lower limit of its adjusted sequence ratio exceeds 1 and picked up by gradient boosting classifier if its probability is greater than 0.5. ADR signals of higher confidence are assigned higher adjusted sequence ratios (rightward) by SSA and higher probabilities (upward) by gradient boosting classifier. It can be observed from Fig. 5 that most known ADRs and unknown potential ADR signals

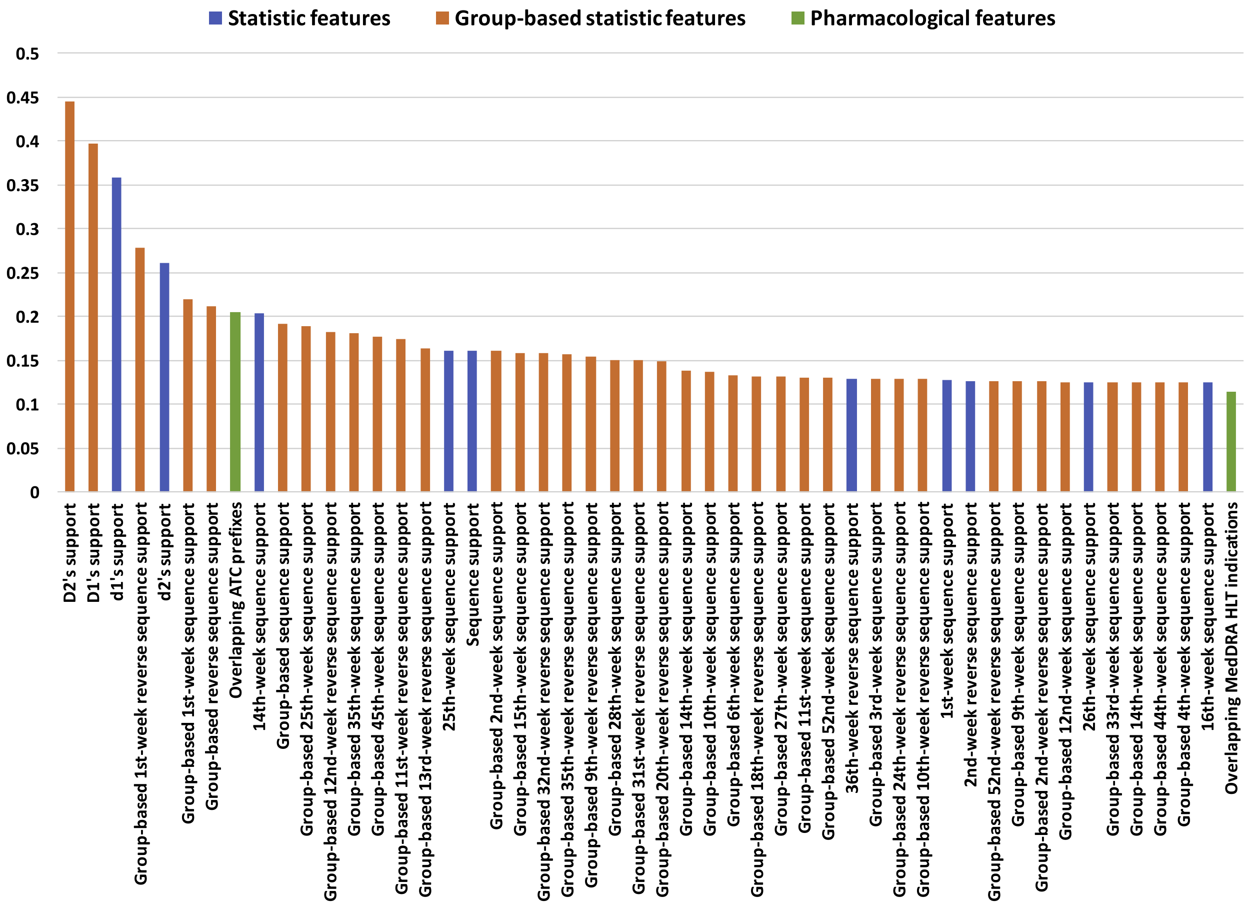


Fig. 4. Top 50 features with highest relative importance in predicting potential ADR signals by gradient boosting classifier.

Table 8
Comparison of known ADRs detected by SSA and gradient boosting classifier.

		Predicted as potential signals by SSA?		
		Yes	No	Total
Predicted as potential signals by gradient boosting classifier?	Yes	31	7	38
	No	1	2	3
	Total	32	9	41

Table 9
Comparison of unknown potential ADR signals detected by SSA and gradient boosting classifier.

		Predicted as potential signals by SSA?		
		Yes	No	Total
Predicted as potential signals by gradient boosting classifier?	Yes	50	10	60
	No	3	2	5
	Total	53	12	65

detected by both SSA and gradient boosting classifier are located in the top left corner, i.e., higher probabilities and lower adjusted sequence ratio. This demonstrates that gradient boosting classifier is more confident about these signals than SSA although previous studies [10–12,35–55] show that it is empirically rare for a signal to have an adjusted sequence ratio above 2.

In addition, Tables 8 and 9 present more detailed comparisons. Table 8 compares the known ADRs detected by gradient boosting classifier and SSA. Among 41 known ADRs, gradient boosting classifier and SSA both detect the 31 same known ADRs (76%). Gradient boosting classifier is able to identify 7 known ADRs that are not detected by SSA (17%). On the other hand, gradient boosting classifier fails to detect 1 known ADR that is picked up by SSA (2%). Both methods are not able to identify 2 known ADRs (5%). Table 9 compares the unknown potential ADR signals detected by gradient boosting classifier and SSA. Both gradient boosting classifier and SSA identify 50/65 same signals (77%). Gradient boosting classifier detects 10 unknown potential ADR signals that are not picked up by SSA (15%). In contrast, gradient boosting classifier does not pick up 3 signals detected by SSA (5%). Both methods consider 12 signals as not potential ADR signals. These results show that gradient boosting classifier not only identifies more known ADRs (15%) but also more unknown potential ADR signals (10%) than SSA.

5. Discussion

It can be observed from the results that gradient boosting classifier has a higher sensitivity than SSA, i.e., is able to detect more known ADRs as well as unknown potential ADR signals. We have also shown that gradient boosting classifier has a comparable specificity with SSA, which means gradient boosting classifier does not pick up more spurious signals than SSA. These results suggest that gradient boosting classifier is a promising signal detection method for ADRs to complement SSA using medication dispensing data. Besides, Wahab et al. [54] found that SSA could detect ADR signals such as the association between rofecoxib and heart attack earlier than spontaneous reports. Thus, gradient boosting classifier has the potential to enhance the timeliness of safety signal detection which will reduce harm and translate to improved health outcomes.

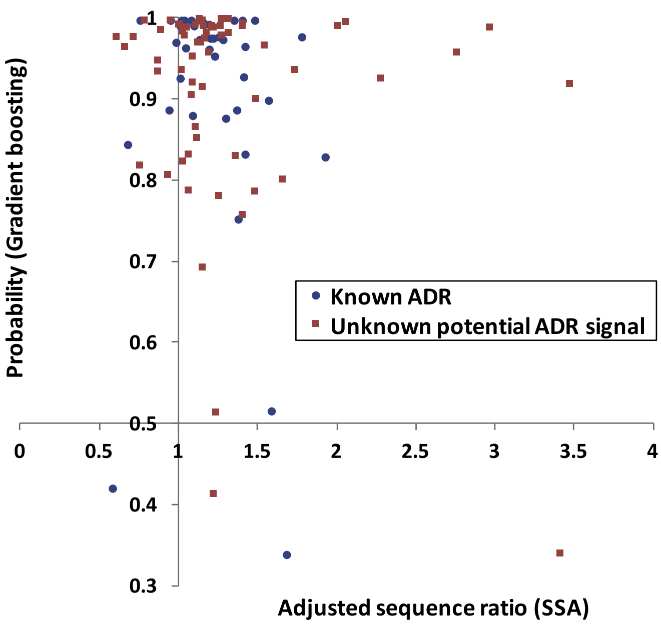


Fig. 5. Comparison of ADR signals detected by SSA (95% confidence interval lower limit of adjusted sequence ratio >1) and gradient boosting classifier (probability >0.5). ADR signals of higher confidence are assigned higher adjusted sequence ratios (rightward) by SSA and higher probabilities (upward) by gradient boosting classifier.

Random forests classifier has been commonly utilized in other settings due to its high predictive performance and very few parameters to tune [22]. In fact, random forests classifier was used in previous studies [20,21] to predict potential ADR signals from the health improvement network (THIN) data in the United Kingdom. One empirical study found that gradient boosting classifier performs better than random forest classifier in eight different metrics [57]. Another empirical study shows that gradient boosting classifier has the best performance among all when the number of features is below 4000 [58]. Our results confirm these observations by showing that gradient boosting classifier outperforms random forests classifier in balancing between all four metrics when there are 247 features.

Supervised ADR classifiers in general and gradient boosting classifier in particular, however, are subject to certain limitations. First, the performance of gradient boosting classifier depends greatly on the quality of training data. Gradient boosting classifier can well detect ADR signals whose combinations of features approximately match those known ADRs in the training set. As such, the deployment of gradient boosting classifier requires firstly building a high quality training set that covers many different types of known ADRs and non-ADRs, and extensive testing on many additional datasets. Second, gradient boosting classifier is difficult to understand by stakeholders. The basic principles of gradient boosting classifier are based on complicated statistical modeling and often viewed as a black-box. Since signal detection tools would be used and assessed by clinicians or medical experts rather than mathematicians, gradient boosting classifier will have lower acceptability than SSA or self-controlled designs whose mechanisms are more straightforward to understand [13]. Another limitation of gradient boosting classifier is the lack of interpretability of results. While each potential ADR signal detected by gradient boosting classifier is associated with a probability, this does not correspond to a risk estimate, which is often expected by clinicians to determine which signals should be further investigated [13].

One way to address these limitations in future work is to combine gradient boosting classifier and SSA in a complementary manner for ADR signal detection. The group of ADR signals detected by both methods should be investigated first, followed by those identified by one method but not picked up by another. The adjusted sequence ratio provided by SSA can then be utilized as a risk estimate to prioritize the ADR signals in each group. Moreover, the output from SSA can be used as a reference to guide and alleviate the dependency of gradient boosting classifier on the training data.

Furthermore, the efficiency of training gradient boosting classifier can be improved to facilitate routine signal detection. At the moment, gradient boosting classifier needs to be re-trained on the whole training data to take into account additional training examples. This process will take longer time as the training data grows bigger, which affects the signal detection. Online gradient boosting [59] could potentially solve this problem by learning new training examples in an incremental manner.

Lastly, a major limitation of using medication dispensing data for ADR signal detection is that only ADRs whose adverse events are treated by medications can be picked up. This limitation can be potentially addressed by relying on additional sources such as unstructured data to provide adverse events. Recent studies have demonstrated the feasibility of detecting ADR signals from unstructured data sources. Wang et al. [60] proposed a method to detect ADRs from clinical notes while White et al. [61] investigated the ADR detection from search log data.

6. Conclusion

ADRs have been creating substantial burden for patients and healthcare systems. Thus, early detection of ADRs could reduce harm and improve peoples health outcomes. In this study, we have demonstrated the utility of SML as a signal detection tool for ADRs in medication dispensing data. We found that gradient boosting classifier achieves the best performance among all the supervised ADR classifiers. In addition, gradient boosting classifier has higher sensi- tivity and comparable specificity in comparison with SSA, a current signal detection method in medication dispensing data. Thus, gradient boosting classifier is a promising ADR signal detection tool to complement SSA.

Acknowledgments

This research has been supported by the Australian Research Council (ARC) DP140103617 and the National Health and Medical Research Council (NHMRC) APP1040938 and APP1110139.

Appendix

Gradient boosting ADR classifier

In this section, we provide some further details of how to learn a gradient boosting ADR classifier. Formally, let $h(x;\alpha_k)$ represent the weak ADR classifier added at iteration k, where α_k is the set of parameters. The gradient boosting ADR classifier represented by $f(x)$ is the weighted combination of weak classifiers added over K iterations.

$$f(x) = X_{k=1}^{K} \beta_k h(x;\alpha_k) \tag{2}$$

where β_k denotes the weight of classifier added at iteration k.

As a result, the goal of learning the gradient boosting ADR classifier is to estimate the parameters $\{\alpha^*, \beta^*\}^K$ such that:

$$\{\alpha^*_k, \beta^*_k\}_{k=1}^{K} = \underset{\{\alpha_k,\beta_k\}_{k=1}^{K}}{\text{argmin}} X_{i=1}^{N} L(y_i, f(x_i)) = \underset{\{\alpha_k,\beta_k\}_{k=1}^{K}}{\text{argmin}} X_{i=1}^{N} L y_i,\ X_{i=1\ \beta_k}^{N} K\, h(x_i, \alpha_k) \tag{3}$$

The process of estimating the parameters is summarized in Algorithm 1 [24]. There are two main steps in each iteration k.

Algorithm 1 Learning algorithm for gradient boosting ADR classifier.

```
1 Initialize $f^{(0)}(x)$
2 for $k = 1$ to $K$ do
3     for $i = 1$ to $N$ do
4         Compute $r_i$ as the negative gradient of the loss function $L\left(y_i, f^{(k-1)}(x_i)\right)$ with respect to $f^{(k-1)}(x_i)$
5     Estimate $\alpha_k$ by fitting $h(x; \alpha)$ to $r$ with least-squares regression
6     Estimate $\beta_k$ by minimizing $\sum_{i=1}^{N} L\left(y_i, f^{(k-1)}(x_i) + \beta_k h(x_i; \alpha_k)\right)$
7     Update $f^{(k)}(x) = f^{(k-1)}(x) + \beta_k h(x; \alpha_k)$
8 return $\hat{f}(x) = f^{(K)}(x)$
```

First, the algorithm estimates the parameters α_k by fitting the weak ADR classifier $h(x;\ \alpha_k\ )$ to the negative gradient of the loss function using least-squares regression (lines 3–5). The insight behind this step is to let the weak ADR classifier correct the prediction errors made in the previous iterations. Then in the second step, the algorithm determines the optimal value of parameters β_k by minimizing the objective loss function specified in Eq. (3). In this study, we empirically set the maximum number of iterations $K = 1000$, employ decision tree as our base weak ADR classifier, and adopt deviance loss function.

References

[1] I.R. Edwards, J.K. Aronson, Adverse drug reactions: definitions, diagnosis, and management, Lancet 356 (9237) (2000) 1255–1259.

[2] M. Drini, Peptic ulcer disease and non-steroidal anti-inflammatory drugs, Aust. Prescr. 40 (3) (2017) 91.

[3] M.-L. Yeh, Y.-J. Chang, P.-Y. Wang, Y.-C.J. Li, C.-Y. Hsu, Physicians responses to computerized drug–drug interaction alerts for outpatients, Comput. Methods Programs Biomed. 111 (1) (2013) 17–25.

[4] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin. Pharmacol. Ther. 91 (6) (2012) 1010–1021.

[5] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Text and data mining techniques in adverse drug reaction detection, ACM Comput. Surv. (CSUR) 47 (4) (2015) 56.

[6] L. Hazell, S.A. Shakir, Under-reporting of adverse drug reactions, Drug Saf. 29 (5) (2006) 385–396.

[7] V.G. Koutkias, M.-C. Jaulent, Computational approaches for pharmacovigilance signal detection: toward integrated and semantically- enriched frameworks, Drug Saf. 38 (3) (2015) 219–232.

[8] I.A. Wahab, N.L. Pratt, L.K. Ellett, E.E. Roughead, Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database, Drug Saf. 39 (4) (2016) 347–354.

[9] L. Mellish, E.A. Karanges, M.J. Litchfield, A.L. Schaffer, B. Blanch, B.J. Daniels, A. Segrave, S.-A. Pearson, The australian pharmaceutical benefits scheme data collection: a practical guide for researchers, BMC Res. Notes 8 (1) (2015) 634.

[10] I.A. Wahab, N.L. Pratt, M.D. Wiese, L.M. Kalisch, E.E. Roughead, The validity of sequence symmetry analysis (ssa) for adverse drug reaction signal detection, Pharmacoepidemiol Drug Saf. 22 (5) (2013) 496–502.

[11] H. Petri, H. De Vet, J. Naus, J. Urquhart, Prescription sequence analysis: a new and fast method for assessing certain adverse reactions of prescription drugs in large populations, Stat. Med. 7 (11) (1988) 1171–1175.

[12] J. Hallas, Evidence of depression provoked by cardiovascular medication: a prescription sequence symmetry analysis, Epidemiology (1996) 478–484.

[13] M. Arnaud, B. Beǵaud, N. Thurin, N. Moore, A. Pariente, F. Salvo, Methods for safety signal detection in healthcare databases: a literature review, Expert Opin. Drug Saf. 16 (6) (2017) 721–732.

[14] E.C.-C. Lai, N. Pratt, C.-Y. Hsieh, S.-J. Lin, A. Pottegård, E.E. Roughead, Y.-H.K. Yang, J. Hallas, Sequence symmetry analysis in pharmacovigilance and pharmacoepidemiologic studies, Eur. J. Epidemiol. (2017) 1–16.

[15] K.P. Murphy, Machine learning: a Probabilistic Perspective, MIT press, Cambridge, MA, USA, 2012.

[16] R. Harpaz, D. Odgers, G. Gaskin, W. DuMouchel, R. Winnenburg, O. Bodenreider, A. Ripple, A. Szarfman, A. Sorbello, E. Horvitz, et al., A time-indexed reference standard of adverse drug reactions, Sci. Data 1 (2014) 140043.

[17] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, Drugbank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (suppl 1) (2006) D668–D672.

[18] P. Mozzicato, Meddra: an overview of the medical dictionary for regulatory activities, Pharm. Med. 23 (2) (2009) 65–75.

[19] W.H. Organization, et al., Who Collaborating Centre For Drug Statistics Methodology: Atc Classification Index With Ddds And Guidelines For Atc Classification And Ddd Assignment, Norwegian Institute of Public Health, Oslo, Norway, 2004.

[20] J.M. Reps, J.M. Garibaldi, U. Aickelin, D. Soria, J.E. Gibson, R.B. Hubbard, Signalling paediatric side effects using an ensemble of simple study designs, Drug Saf. 37 (3) (2014) 163–170.

[21] J.M. Reps, J.M. Garibaldi, U. Aickelin, J.E. Gibson, R.B. Hubbard, A supervised adverse drug reaction signalling framework imitating bradford hills causality considerations, J. Biomed. Inform. 56 (2015) 356–368.

[22] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[23] H. Jin, J. Chen, H. He, C. Kelman, D. McAullay, C.M. O'Keefe, Signaling potential adverse drug reactions from administrative health databases, IEEE Trans. Knowl. Data Eng. 22 (6) (2010) 839–853.

[24] L. Guelman, Gradient boosting trees for auto insurance loss cost modeling and prediction, Expert Syst. Appl. 39 (3) (2012) 3659–3667.

[26] C.E. McCulloch, Generalized linear models, J. Am. Statist. Assoc. 95 (452) (2000) 1320–1324.

[25] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[27] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[28] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1) (1990) 109–118.

[33] P. Bühlmann, T. Hothorn, Boosting algorithms: regularization, prediction and model fitting, Stat. Sci. (2007) 477–505.

[30] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[32] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, Boca Raton, FL, USA, 1984.

[29] G. Bebis, M. Georgiopoulos, Feed-forward neural networks, IEEE Potentials 13 (4) (1994) 27–31.

[34] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd International Conference On Machine Learning, ACM, 2006, pp. 233–240.

[35] I. Tsiropoulos, M. Andersen, J. Hallas, Adverse events with use of antiepileptic drugs: a prescription and event symmetry analysis, Pharmacoepidemiol. Drug Saf. 18 (6) (2009) 483–491.

[36] G. Lindberg, J. Hallas, Cholesterol-lowering drugs and antidepressants–a study of prescription symmetry, Pharmacoepidemiol Drug Saf. 7 (6) (1998) 399–402.

[37] D.J. Cher, Myocardial infarction and acute cholecystitis: an application of sequence symmetry analysis, Epidemiology 11 (4) (2000) 446–449.

[38] P. Bytzer, J. Hallas, Drug-induced symptoms of functional dyspepsia and nausea. A symmetry analysis of one million prescriptions, Aliment. Pharmacol. Ther. 14 (11) (2000) 1479–1484.

[39] G. Corrao, E. Botteri, V. Bagnardi, A. Zambon, A. Carobbio, C. Falcone, O. Leoni, Generating signals of drug-adverse effects from prescrip- tion databases and application to the risk of arrhythmia associated with antibacterials, Pharmacoepidemiol Drug Saf. 14 (1) (2005) 31–40.

[40] E.L. Thacker, S. Schneeweiss, Initiation of acetylcholinesterase inhibitors and complications of chronic airways disorders in elderly patients, Drug Saf. 29 (11) (2006) 1077–1085.

[41] L. Silwer, M. Petzold, J. Hallas, C.S. Lundborg, Statins and nonsteroidal anti-inflammatory drugsan analysis of prescription symmetry, Pharmacoepidemiol Drug Saf. 15 (7) (2006) 510–511.

[42] S. Vegter, D. Jong-van den Berg, T. Lolkje, Misdiagnosis and mistreatment of a common side-effect–angiotensin-converting enzyme inhibitor-induced cough, Br. J. Clin. Pharmacol. 69 (2) (2010) 200–203.

[43] S. Vegter, P. de Boer, K.W. van Dijk, S. Visser, et al., The effects of antitussive treatment of ace inhibitor-induced cough on therapy compliance: a prescription sequence symmetry analysis, Drug Saf. 36 (6) (2013) 435–439.

[44] K.B. Pouwels, S.T. Visser, H.J. Bos, E. Hak, Angiotensin-converting enzyme inhibitor treatment and the development of urinary tract infections: a prescription sequence symmetry analysis, Drug Saf. 36 (11) (2013) 1079–1086.

[45] J.F. van Boven, S. Vegter, et al., Inhaled corticosteroids and the occurrence of oral candidiasis: a prescription sequence symmetry analysis, Drug Saf. 36 (4) (2013) 231–236.

[46] L.M. Kalisch Ellett, N.L. Pratt, J.D. Barratt, D. Rowett, E.E. Roughead, Risk of medication-associated initiation of oxybutynin in elderly men and women, J. Am. Geriatr. Soc. 62 (4) (2014) 690–695.

[47] M. Takada, M. Fujimoto, K. Yamazaki, M. Takamoto, K. Hosomi, Association of statin use with sleep disturbances: data mining of a spontaneous reporting database and a prescription database, Drug Saf. 37 (6) (2014) 421–431.

[48] L. Rasmussen, J. Hallas, K.G. Madsen, A. Pottegård, Cardiovascular drugs and erectile dysfunction–a symmetry analysis, Br. J. Clin. Pharmacol. 80 (5) (2015) 1219–1223.

[49] Y. Takeuchi, K. Kajiyama, C. Ishiguro, Y. Uyama, Atypical antipsychotics and the risk of hyperlipidemia: a sequence symmetry analysis, Drug Saf. 38 (7) (2015) 641–650.

[50] K.B. Pouwels, N.N. Widyakusuma, J.H. Bos, E. Hak, Association between statins and infections among patients with diabetes: a cohort and prescription sequence symmetry analysis, Pharmacoepidemiol Drug Saf. 25 (10) (2016) 1124–1130.

[51] G.E. Caughey, E.E. Roughead, N. Pratt, S. Shakib, A.I. Vitry, A.L. Gilbert, Increased risk of hip fracture in the elderly associated with prochlorperazine: is a prescribing cascade contributing? Pharmacoepidemiol Drug Saf. 19 (9) (2010) 977–982.

[52] J.A. Cole, F.A. Farraye, H.J. Cabral, Y. Zhang, K.J. Rothman, Irritable bowel syndrome and hysterectomy: a sequence symmetry analysis, Epidemiology 18 (6) (2007) 837–838.

[53] S.R. Garrison, C.R. Dormuth, R.L. Morrow, G.A. Carney, K.M. Khan, Nocturnal leg cramps and prescription use that precedes them: a sequence symmetry analysis, Arch. Intern. Med. 172 (2) (2012) 120–126.

[54] I.A. Wahab, N.L. Pratt, L.M. Kalisch, E.E. Roughead, Comparing time to adverse drug reaction signals in a spontaneous reporting database and a claims database: a case study of rofecoxib-induced myocardial infarction and rosiglitazone-induced heart failure signals in australia, Drug Saf. 37 (1) (2014) 53–64.

[55] E.E. Roughead, E.W. Chan, N.-K. Choi, M. Kimura, T. Kimura, K. Kubota, E.C.-C. Lai, K.K. Man, T.A. Nguyen, N. Ooba, et al., Variation in association between thiazolidinediones and heart failure across ethnic groups: retrospective analysis of large healthcare claims databases in six countries, Drug Saf. 38 (9) (2015) 823–831.

[56] L. Breiman, et al., Classification and Regression Trees, Chapman & Hall, New York, NY, USA, 1984.

[57] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: W.W. Cohen, A. Moore (Eds.), Proceedings Of The 23rd International Conference On Machine Learning, (Eds.), ACM, New York, NY, USA, 2006, pp. 161–168.

[58] R. Caruana, N. Karampatziakis, A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, in: W.W. Cohen, A. McCallum, S.T. Roweis (Eds.), Proceedings Of The 25th International Conference On Machine Learning, (Eds.), ACM, New York, NY, USA, 2008, pp. 96–103.

[59] A. Beygelzimer, E. Hazan, S. Kale, H. Luo, Online gradient boosting, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Proceedings Of The Advances In Neural Information Processing Systems 28, (Eds.), Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2015, pp. 2458–2466.

[60] G. Wang, K. Jung, R. Winnenburg, N.H. Shah, A method for systematic discovery of adverse drug events from clinical notes, J. Am. Med. Inform. Assoc. 22 (6) (2015) 1196–1204.

[61] R.W. White, S. Wang, A. Pant, R. Harpaz, P. Shukla, W. Sun, W. DuMouchel, E. Horvitz, Early identification of adverse drug reactions from search log data, J. Biomed. Inform. 59 (2016) 42–48.