Artificial Intelligence In Medicine 104 (2020) 101839

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Detecting potential signals of adverse drug events from prescription data

Chen Zhan[a,*], Elizabeth Roughead[b], Lin Liu[a], Nicole Pratt[b], Jiuyong Li[a]

[a] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, South Australia 5095, Australia
[b] School of Pharmacy and Medical Sciences, Quality Use of Medicines and Pharmacy Research Centre, Sansom Institute, University of South Australia, Adelaide, South Australia 5000, Australia

ARTICLE INFO

Keywords:
Adverse drug events (ADEs)
Prescription data
Logistic regression
Case-crossover

ABSTRACT

Adverse drug events (ADEs) may occur and lead to severe consequences for the public, even though clinical trials are conducted in the stage of pre-market. Computational methods are still needed to fulfil the task of pharmacosurveillance. In post-market surveillance, the spontaneous reporting system (SRS) has been widely used to detect suspicious associations between medicines and ADEs. However, the passive mechanism of SRS leads to the hysteresis in ADE detection by SRS based methods, not mentioning the acknowledged problem of under-reporting and duplicate reporting in SRS. Therefore, there is a growing demand for other complementary methods utilising different types of healthcare data to assist with global pharmacosurveillance. Among those data sources, prescription data is of proved usefulness for pharmacosurveillance. However, few works have used prescription data for signalling ADEs. In this paper, we propose a data-driven method to discover medicines that are responsible for a given ADE purely from prescription data. Our method uses a logistic regression model to evaluate the associations between up to hundreds of suspected medicines and an ADE spontaneously and selects the medicines possessing the most significant associations via Lasso regularisation. To prepare data for training the logistic regression model, we adapt the design of the case-crossover study to construct case time and control time windows for the extraction of medicine use information. While the case time window can be readily determined, we propose several criteria to select the suitable control time windows providing the maximum power of comparisons. In order to address confounding situations, we have considered diverse factors in medicine utilisation in terms of the temporal effect of medicine and the frequency of prescription, as well as the individual effect of patients on the occurrence of an ADE. To assess the performance of the proposed method, we conducted a case study with a real-world prescription dataset. Validated by the existing domain knowledge, our method successfully traced a wide range of medicines that are potentially responsible for the ADE. Further experiments were also carried out according to a recognised gold standard, our method achieved a sensitivity of 65.9% and specificity of 96.2%.

1. Introduction

An adverse drug event (ADE) is an "unfavourable medical event that occurs in association with the use of a certain medication" [1]. In the United States, medicine-related adverse events (both morbidity and mortality) were estimated to cost $76.6 billion annually [2], and the U.S. Food and Drug Administration (FDA) receives approximately 0.25 million reports of suspected ADEs per year [3]. As one of the leading concerns of medicine safety issues, ADEs have already raised attention from both regulatory agencies and research communities.

To assess the efficacy and safety of a new medicine, clinical trials are normally undertaken before introducing it to the market. However, clinical trials suffer from insufficient samples and limited durations for ADE detection. Clinical trials may not include all subgroups of patients who eventually will use the medicine in the real world because subpopulations such as elderly, children, pregnant women, patients with multiple diseases are usually excluded from the trials [3]. This limits the capacity of clinical trials to detect infrequent and rare ADEs. The short observation period in a trial may also limit its ability to detect ADEs with a longer latency. Therefore, post-market surveillance is still needed for monitoring ADEs and discovering medicines that are responsible for ADEs.

As an important resource for post-market surveillance, spontaneous reporting system (SRS) has been used by many regulatory authorities such as the adverse event reporting system (AERS) from the U.S. FDA. SRS encourages health professionals to report suspected ADEs along

* Corresponding author.
E-mail addresses: chen.zhan@mymail.unisa.edu.au (C. Zhan), Libby.Roughead@unisa.edu.au (E. Roughead), Lin.Liu@unisa.edu.au (L. Liu), Nicole.Pratt@unisa.edu.au (N. Pratt), Jiuyong.Li@unisa.edu.au (J. Li).

https://doi.org/10.1016/j.artmed.2020.101839
Received 22 February 2019; Received in revised form 6 February 2020; Accepted 24 February 2020

with the specific medicines involved. Many methods have been developed to determine the associations between reported ADEs and suspected medicines using SRS databases. Those methods are mostly based on measures of disproportionality (which quantify the extent to which an ADE is reported disproportionately with a certain medicine in comparison to other medicines from the whole SRS database [4]), such as proportional reporting ratio (PRR) [5], reporting odds ratio (ROR) [6], the information component (IC) used by Bayesian confidence propagation neural network (BCPNN) [7]. In general, those methods project data onto contingency tables to calculate the measures according to different equations and further determine whether an association exists by comparing to a pre-defined threshold.

It is widely acknowledged that SRS suffers from the problems of under-reporting and duplicate reporting [8]. Thus the findings could be biased towards highly reported cases and rare cases tend to be overlooked. Furthermore, the procedure of reporting heavily relies on the suspicion that a medicine has caused a corresponding ADE, which involves subjective judgement and may result in varying quality, inconsistency or incompletion of reports. SRS based methods are passive ADE discovery, which also limit the discovery power of the methods. The procedure of reporting an ADE to SRS, to some extent, is the procedure of generating a hypothesis that the suspected medicine causes the corresponding ADE. So those SRS-based ADE discovery methods are mainly meant to evaluate the hypotheses according to the designed criteria. Such characteristic of SRS leads to the hysteresis in ADE detection while millions of patients have already been exposed [9].

To complement the SRS which suffers from the above-mentioned limitations, there is a growing demand for other methods utilising different types of healthcare data to assist with global pharmacosurveillance. Previous research has shown that the diversity of signalling methods and data sources facilitates more ADEs being detected and can potentially strengthen post-market surveillance of medicine safety [10]. Prescription data is such a data source with proved usefulness for pharmacosurveillance [11]. In contrast to SRS, prescription data does not suffer from the bias of subjective reporting as it faithfully records every dispensation of prescribed medicines irrespective of whether or not an ADE is suspected. Compared to other electronic health records (EHRs), prescription data covers a diverse range of populations and is more accessible [12]. Furthermore, as a longitudinal observational data source, prescription data facilitates data-driven ADE discovery that automatically generates and evaluates the hypotheses from data regarding ADEs without human involvement. Therefore, some previously ignored or unknown ADEs may be discovered [13].

There is only a handful of methods mainly relying on prescription datasets containing no additional information for ADE discovery, due to the uncertainty regarding the outcomes of medicines. As a prescription dataset does not explicitly record any occurrence of ADEs, it is challenging to signal ADEs from a prescription dataset. The prescription sequence symmetry analysis (PSSA), which was introduced in 1988 and further developed by Petri et al. [14] and Hallas [15], is one of the few methods which prescription data. To impute the outcomes of medicines, PSSA uses specific medicines for treating ADEs as indicators of corresponding ADEs. For example, the dispensation of doxycycline, an antibiotic medicine, indicates that the patient is most likely to suffer from a bacterial infection while the type of infection is unknown. Some case studies have been conducted with PSSA [16–18], but they were all done by pairs of medicine-event (examining a medicine with an event at one time). However, due to its low cost, effectiveness and simplicity, PSSA is still an important tool for pharmacosurveillance.

In this paper, we aim to detect medicines that are responsible for a given ADE from pure prescription data. For this purpose, we develop a logistic regression based method to evaluate the associations between multiple medicines and the ADE spontaneously. To prepare data for training the logistic regression model, we follow the design of the case-crossover study [19] to construct case time (immediately before an ADE occurs) and control time (when no ADE occurs) windows for the extraction of medicine use information. Therefore, the confounding situation raised by fixed characteristics of patients, such as gender can be eliminated. While the case time window can be easily determined, we propose several criteria to select the suitable control time windows and ensure that medicine use information in the control time windows can be effectively compared with that in the case time windows. In practice, a considerable number of candidate medicines are taken by patients subject to a specific ADE. Most of the candidate medicines do not have an obvious effect on the ADE of interest while only a few do. Hence, we provide a sparse solution to select the candidate medicines with the most significant associations via Lasso (least absolute shrinkage and selection operator) [20] regularisation. While the basic logistic regression uses binary indicators which can only distinguish whether a medicine is used or not, thereby fails to address the additional confounding situations, we further propose a weight function to consider the various factors in medicine utilisation, including the temporal effect and the frequency of prescription. We also consider the individual effect of patients on the occurrence of the ADE to eliminate the unrelated medicines from the candidates.

To assess the performance of the proposed method, we conducted a case study using frusemide initiation as the indicator of the ADE (heart/cardiac failure and/or peripheral oedema) with a real-world prescription dataset from the Pharmaceutical Benefits Scheme of Australia. Validated by the existing database of product information, our method successfully traced a wide range of medicines that are potentially responsible for the ADE with a precision between 50% and 80%. Further experiments were also carried out according to a recognised gold standard of 20 ADEs along with 41 positive and 53 negative medicines which can cause the ADEs or unlikely to cause the ADEs respectively, our method achieved a sensitivity of 65.9% and a specificity of 96.2%.

The major novelty of this paper is summarised as follows:

- Given an ADE of interest, our method specifically focuses on spontaneous signalling as many ADE-causing medicines as possible. As a complement to other signal detection methods for ADEs, our work can potentially strengthen the global pharmacosurveillance.
- In order to overcome the confounding situation raised by time-invariant characteristics of patients, we use the design of the case-crossover study to extract medicine use information for data preparation.
- While the basic logistic regression uses binary indicators which can only distinguish whether a medicine is used or not, our method considers the various factors of medicine utilisation, including the temporal effect of a medicine and the frequency of a prescription, as well as the individual effect of patients on the occurrence of an ADE.

The rest of this paper is organised as follows. In Section 2, we introduce the necessary background regarding the case-crossover study and define the research problem. Section 3 presents the proposed method in details, including data preparation, the basic logistic regression model with a sparse solution and the two improvements to it. Section 4 introduces the settings of experiments and presents the results. In Section 5, we discuss some interesting findings in relation to the developed method. Finally, Section 6 concludes the paper and suggests future work.

2. Problem definition

2.1. Case-crossover study

Randomised controlled trials (RCTs) are the widely acknowledged golden standard for causal inference. However, RCTs are not always feasible because of the cost and ethical concerns; e.g. it will be unethical to require an experiment participant to expose himself/herself to risk factors which could be life-threatening. Therefore, various methods have been developed for inferring strong associations or possible causal relationships between an exposure and the outcome from

observational data. It has been shown that well-designed observational studies can achieve comparative results to those of RCTs [21].

The case–control study [22] is a commonly used type of retrospective observational study. In a case-control study, two groups are defined on the basis of observational data, one including the cases, i.e. samples with the outcome of interest (e.g. lung cancer) and the other including the controls, i.e. samples without the outcome. By assuming that the exposure between the two groups can represent each other if their outcome states were exchanged, then a significant difference between the two groups in the level of exposure to a risk factor (e.g. smoking) will suggest a strong association between the risk factor and the outcome.

A case-crossover study [19] could be considered a crossover version of a case–control study, where each case serves as its own control. In a case-crossover study, all the subjects who experienced the outcome (event) of interest are cases and thus are included in the study. Instead of conducting comparisons between two different groups, the case-crossover study compares the difference in the levels of exposure to the risk factor of the cases from the same group but during different time windows, i.e. case time windows and control time windows. Hence, the case-crossover study is immune to a selection bias. The case-crossover study was introduced by Maclure [19] to answer the question "Was this event triggered by something unusual that happened just before?" [23]. The assumption is that if the risk factor potentially causes the outcome, the risk factor should appear, at least, more frequently immediately prior to the occurrence of the outcome (case time windows) than at any similar period distant from the occurrence of the outcome (control time windows). To identify such associations given a specific ADE, the concept of case-crossover study is utilised in our method to prepare a raw prescription dataset for a logistic regression model.

2.2. Research problem

In epidemiology, given that the other conditions are fixed, if an ADE "would not occur at all or would not occur until some later time" [24] without the antecedent use of a medicine, then the ADE might be caused by the specific medicine. Therefore, the proposed research problem addressed in this paper can be defined as follows.

Research Problem. For a given ADE of interest and a prescription dataset containing dispensing records of prescription with timestamps, we aim to find such associations between a medicine and the ADE that satisfy **Condition (1)** the medicine was dispensed prior to and not distant from the occurrence of the ADE, **Condition (2)** a dispensation of the medicine is frequently co-occurred with the ADE, and **Condition (3)** when the ADE does not occur, the medicine is most likely absent.

3. Method

In this section, we will first present the data preparation process from the prescription dataset according to the design of a case-crossover study. Then, we will introduce the basic logistic regression model with Lasso regularisation. Furthermore, we propose two improvements to the basic model by taking more factors into account.

3.1. Data preparation

3.1.1. Case and control time windows on ADE

In the following text, according to the design of a case-crossover study [19], we propose the definition of the case and the control time windows in the context of discovering medicines that might lead to a given ADE.

Let $\mathcal{E}$ represent a given ADE of interest. We use y to indicate whether $\mathcal{E}$ occurs ($y = 1$) or not ($y = 0$). In the prescription dataset, the occurrence of $\mathcal{E}$ could be imputed on the basis of the prescription of the specific medicines used to treat $\mathcal{E}$. We call these medicines *ADE-indicator* medicines in this paper. From SIDER [25], a well-known database for marketed medicines and their recorded adverse reactions, we collect the list of ADE-indicator medicines for $\mathcal{E}$. The temporal ordering of a medicine suspected to have caused the ADE, the ADE and the ADE-indicator medicine is illustrated in Fig. 1. The ADE occurs after the dispensation of a suspected medicine and the ADE-indicator medicine is dispensed shortly after the occurrence of the ADE as its treatment. While the occurrence of an ADE is unobserved in the prescription dataset, the dispensation of an ADE-indicator medicine can be used as the indication of the occurrence of the ADE. Here, we only use the initial prescription of ADE-indicator medicine as the proxy of the occurrence of $\mathcal{E}$. The initial prescription reflects the status of the patients accurately with less interference (while the following prescriptions could be attributed to other reasons, e.g. resupply of the previous prescription).

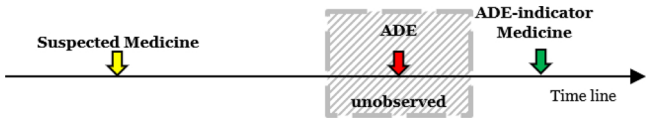

Fig. 1. Temporal ordering of suspected medicine, ADE and ADE-indicator medicine.

Based on the principle that a cause happens prior to its consequence [26], the medicines responsible for $\mathcal{E}$ must be dispensed within a time window that is immediately before the occurrence of $\mathcal{E}$. Therefore, for the case-crossover study in the context of ADEs, we define the case time window as follows.

Definition 1. Case Time Window: Given an ADE of interest, $\mathcal{E}$, the case time window for a patient is a period of τ days immediately before the day when $\mathcal{E}$ occurs to the patient. Here, τ is the pre-defined length of time (in days) based on the maximum induction time of the prescribed medicine during which the medicine keeps exerting its effect on patients.

Note that in this paper, we only use the initial prescription of ADE-indicator medicine to impute the occurrence of the ADE to the patient, this ensures there is one unique case time window specific to each patient and avoids possible conflicts during constructing different time windows. While the case time windows contain prescription records of the suspected medicines potentially responsible for the occurrence of $\mathcal{E}$, control time windows are needed to provide a comparison for verifying the suspected medicines and eliminating the unrelated medicines.

Definition 2. Control Time Window: Given the case time window constructed for a patient according to the occurrence of the ADE $\mathcal{E}$, his/her control time window is the time window prior to and not overlapping his/her case time window. Furthermore, during the control time window, the patient is not affected by $\mathcal{E}$.

To achieve the maximum statistical power of the logistic regression model and prevent the possible statistical bias towards the oversized sample, we construct the same number of control time windows as that of the case time windows for each patient in the analysis (i.e. one case time window versus one control time window for one patient).

The two types of time windows contain the information regarding medication use, such as which medicine was dispensed, the number of prescriptions with this medicine, and the dates of these dispensations. We refer to these as the *medicine use information* in the rest of this paper. The case time window is constructed immediately before the occurrence of $\mathcal{E}$; therefore, it corresponds to the outcome that $y = 1$. Similarly, the control time window corresponds to $y = 0$ as no ADE occurs when the control time window ends.

3.1.2. Candidate medicines

By locating the initial prescription of the ADE-indictor medicines in the dataset, we could include all patients who have experienced $\mathcal{E}$ in the analysis and index them by $i \in \{1, ..., N\}$. For each patient, the case

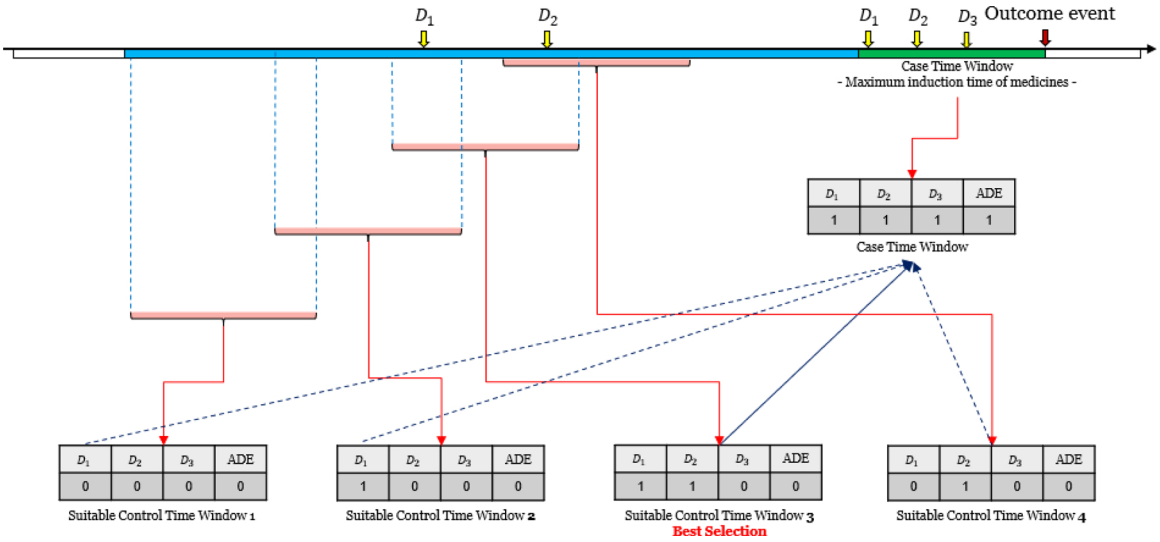

Fig. 2. Example of selecting the best control time window for a case time window. (For interpretation of the references to colour in this figure citation, the reader is referred to the web version of this article.)

time window could be readily determined according to Definition 1. We presume that all the medicines dispensed within the case time windows are suspected of causing $\mathcal{E}$, and any medicines dispensed outside case time window are unrelated to $\mathcal{E}$ (**Condition 1** in the definition of **Research Problem**). Hence, as defined below, we obtain the set of candidate medicines that might lead to the ADE $\mathcal{E}$.

Definition 3. Set of Candidate Medicines: A candidate medicine (i.e. a risk factor for the case-crossover study) is any medicine that is dispensed to a patient in the analysis during his case time window. Let $\mathbf{D}_i$ denote the set of medicines dispensed during the case time window of patient i in the prescription dataset, then the set of all candidate medicines, denoted as $\mathbf{D}^*$, is defined as $\mathbf{D}^* = \bigcup_i \mathbf{D}_i$.

3.1.3. Selecting control time windows

In the case-crossover study, for a given patient, the exposure state (medicine use information) in his/her control time window should impute that in his/her case time window if the outcome ($\mathcal{E}$) is removed. To ensure this, the control time window needs to be carefully selected in terms of both the length and the starting point. Hence, specific to signalling potential ADEs, we propose the following criteria for selecting the appropriate control time windows:

1. **The control time window should have an equal length as the case time window, i.e. τ days for every patient in the analysis.** This is the fundamental criterion to eliminate bias during sampling medicine use information. The equal length of the control time window and the case time window ensures a fair comparison that a given medicine has the same probability of being dispensed in the control time window and in the case time window if it is unrelated to $\mathcal{E}$.
2. **The control time window should not overlap with the case time window.** During the case time window, the patient is exposed to the suspected medicines that might be responsible for the given ADE. As the control time window is selected to provide comparisons to the case time window, the control time window cannot be contaminated by exposures in the case time window. Hence, the control time window should not overlap with the case time window.
3. **The control time window should be constructed sufficiently close to its case time window.** The case-crossover study is immune to the selection bias as each patient in the analysis serves as his/her own control. Hence, it eliminates the confounding factors which are the patient characteristics that remain constant over time. However, as some of the patient characteristics might change over time, which may introduce certain confounding factors. To eliminate (or minimise) such confounding factors, the control time window should be constructed sufficiently close to its case time window so that the change in the patient characteristics is negligible.
4. **The control time window which contains the maximum medicine use information of candidate medicines should be chosen.** In practice, there might be several suitable time windows that can be chosen as the control time window for a case time window. These suitable time windows contain different numbers of records regarding candidate medicine use. Naturally, a time window during which the patient did not take any medicine and thus no ADE occurred can be assigned as the control time window. Discovering medicines that are potentially responsible for the ADE requires us to efficiently rule out medicines that are unrelated to the ADE. If we choose such a control time window containing no dispensations of medicines, then we would not have the information to eliminate the unrelated medicines from the set of candidate medicines. To sample as much medicine use information in the control time window as possible, we choose the control time window that contains the maximum use information of the candidate medicines from all the suitable time windows.

We use the example shown in Fig. 2 to demonstrate all of the above criteria. Suppose that a patient experienced the ADE of interest, as the red arrow above the timeline shows, and a case time window could be constructed by tracing back τ days (the maximum induction time of medicines) immediately before the ADE occurs, as shown by the green area along the timeline. In the case time window, three medicines (D_1, D_2 and D_3) were dispensed and hence are suspected of leading to the ADE. Before the ADE occurred, the patient was dispensed with D_1 and D_2. This suggested that D_1 and D_2 are less suspicious for leading to the ADE. The blue area is the range of pre-defined "sufficient closeness" to the occurrence of the event. Therefore, the control time window should be selected from the range of "sufficient closeness" and any time window beyond this range should not be selected as control. The four time windows of equal length as the case time window were indicated, and all of them could be chosen as the control time window based on Definition 2. However, the third time window was the best selection, as it captured more prescriptions than the others. It is also worth mentioning that, if there are multiple time windows containing the

maximum medicine use information within the range of "sufficient closeness", we should choose the one closest to the ADE.

3.1.4. Extracting medicine use information

Instinctively, a medicine frequently appears in the case time windows (frequently co-occurs with the ADE) is suspicious of causing the ADE (**Condition 2** in the definition of **Research Problem**). While no ADE occurs (during control time windows), the suspected medicine is most likely to be absent (**Condition 3** in the definition of **Research Problem**). If a medicine violates **Condition 3**, i.e. being dispensed in the control time windows, then the suspicion on it becomes questionable as its presence leads to no ADE. To determine which medicines in the candidate set are potentially responsible for the ADE, we need to extract the use information of the candidate medicines from both the case time windows and control time windows. Therefore, the use information of these medicines can be compared and analysed.

We use indicator $d_m \in \{0, 1\}$ to represent whether a candidate medicine $D_m \in \mathbf{D}^*$ was used in a specific time window ($d_m = 1$) or not ($d_m = 0$). Thus, we obtain a vector of indicators $\mathbf{V} = \{d_1, ..., d_m, ..., d_M\}$ for each time window, where $M = |\mathbf{D}^*|$ is the total number of candidate medicines.

The flow chart in Fig. 3 summarises the steps of the data preparation process.

3.2. Logistic regression model with Lasso regularisation

3.2.1. Basic logistic regression model

Considering that there are only binary outcomes in our analysis, i.e. $y = 0$ (no ADE occurs immediately after the control time windows) or $y = 1$ (ADE occurs immediately after case time windows), the logistic regression model [27] could be used to estimate the effects of the medicines on the ADE of interest.

To estimate the effects of the candidate medicines, we assign a coefficient β_m to each medicine D_m in the candidate set $\mathbf{D}^*$ accordingly. Thus, given the vector $\mathbf{V} = \{d_1, ..., d_m, ..., d_M\}$ of the medicine use information from $\mathbf{D}^*$, we represent the probability of outcome y given the vector $\mathbf{V}$ regarding the medicine use information as follows:

$$\begin{aligned} \Pr(y_k|\mathbf{V}_k) &= f(z) \\ &= \frac{1}{1 + e^{-(\sum_{m=1}^{M} \beta_m d_m + \varepsilon)}} \end{aligned} \tag{1}$$

where k indexes the time windows (case and control time windows) that were constructed for all the patients in the analysis and K is the total number of time windows.

In this setting, fitting the logistic regression model is equivalent to solving the following optimisation problem:

$$\underset{\vec{\beta}}{\operatorname{argmin}} = -\frac{1}{K} \sum_{k=1}^{K} [y_k * \mathbf{V}_k \vec{\beta} - \log(1 + e^{(\mathbf{V}_k \vec{\beta})})] \tag{2}$$

where $\vec{\beta} = \{\beta_1, \beta_2, ..., \beta_M\}$ is the vector of the effect parameters.

In this equation, the parameters $\beta_m | m \in \{1, ..., M\}$ are of interest. They can be interpreted as the effects of the corresponding candidate medicine $D_m \in \mathbf{D}^*$ on ADE $\mathcal{E}$. In [28], such "effects" were explained as the "general capacities to transmit changes among variables" (i.e. the use of candidate medicines and the occurrence of ADE). More specifically, if the coefficient β_m is evaluated to be positive, exposure to medicine D_m would contribute to the occurrence of ADE $\mathcal{E}$ to some extent.

3.2.2. Feature selection via Lasso regularisation

In practice, patients often take several medicines concurrently before an ADE occurs. Considering that a large number of patients are included in the analysis, there could be a considerable number of medicines in the candidate set. This poses a considerable challenge to detect the medicines responsible for the ADE. Suppose there are M candidate medicines; then, we will obtain 2^M possible combinations (subsets) of the medicines. For this large number, it is computationally prohibitive to evaluate each subset to find all the medicines responsible for the ADE. To achieve a sparse result and select medicines that are the most likely to be responsible for the ADE, we introduce Lasso regularisation into the logistic regression as follows:

$$\underset{\vec{\beta}}{\operatorname{argmin}} = -\frac{1}{K} \sum_{k=1}^{K} [y_k * \mathbf{V}_k \vec{\beta} - \log(1 + e^{(\mathbf{V}_k \vec{\beta})})] + \lambda \|\vec{\beta}\|_1 \tag{3}$$

Lasso was first formulated by Robert Tibshirani in 1996 [20]. Lasso not only fulfils the task of model regularisation to prevent overfitting problem but also conducts feature selection. Its reliability has been proved in many studies across different research areas, including pharmacoepidemiology [29]. By constraining the sum of the absolute value of the coefficients, Lasso forces many coefficients to zero and allows only relevant features to have nonzero coefficients. Therefore, most components of $\vec{\beta}$ will be evaluated as zero due to the ignorable associations between these medicines and the ADE. As the output of our method, a small proportion of medicines with non-zero (positive) coefficients will be selected from the set of candidate medicines. We further rank these medicines according to their coefficients β_m in the descending order as the coefficients indicate the strength of the association.

3.3. Considering various factors in medicine utilisation

The variety of medicine utilisation before the ADE occurs results in different levels of exposure to the medicines. Such variety not only

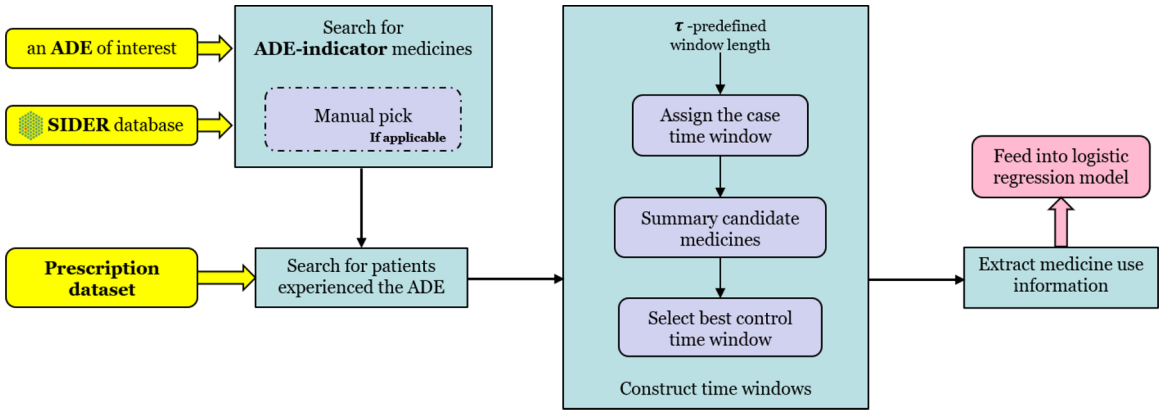

Fig. 3. Flow chart of data preparation.

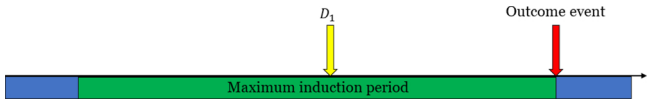

Fig. 4. Common situation wherein medicine is dispensed and the outcome event occurs. (For interpretation of the references to colour in this figure citation, the reader is referred to the web version of this article.)

comes from using different medicines but also from different situations in terms of using the same medicine, i.e. the temporal effect and the frequency of prescription. Considering various factors facilitates addressing additional confounding situations in medicine utilisation, thereby achieving more accurate estimation of the effects of medicines on the ADE. However, in the basic solution, those two kinds of information regarding medicine use is ignored, thus limiting the performance of the method. Therefore, we introduce a function to measure the levels of exposure by weighting the temporal and frequency effects of prescribed medicines.

3.3.1. Temporal effect and frequency of prescription

First, we will explain why the temporal effect and the frequency of a prescription are essential in leading to the ADE with three examples. In the first example, a medicine was dispensed before the ADE of interest occurred, as shown in Fig. 4, where the red arrow indicates the occurrence of the ADE and the green area represents the maximum induction period. The medicineD_1 was dispensed roughly in the middle of the maximum induction period, marked by the yellow arrow. Let us set this as a common situation of medicine use regarding D_1. However, the situations may vary in two ways in general, i.e. the frequency of prescriptions of dispensing D_1 and the time gap between the prescription of dispensing D_1 and the occurrence of the ADE, as illustrated in the following.

During the maximum induction period immediately before the ADE, a patient may be dispensed with several prescriptions of the same medicine. As shown in Fig. 5, within the time window before the occurrence of the ADE, the patient was dispensed with D_1 twice. Unlike the situation in Fig. 4, twice the frequency of prescriptions brings additional exposure to the medicine. Obviously, the frequency of prescriptions reflects the degree of exposure, which results in the different incidences of the ADE, and we refer to it as the effect of frequency of prescription on the ADE.

Furthermore, the time gap between the dispensation of a prescription and the occurrence of ADE might vary. As shown in Fig. 6, the time gap between the prescription and the ADE is longer than that in Fig. 4. According to the previous research regarding the relative risk (RR) of ADEs [30], presuming that the patients will keep taking the medicine after the prescription is dispensed (until all the pills are consumed), the RR fluctuates over time rather than remaining constant. Fig. 7 shows the curve of the RR with respect to the occurrences of the ADE, which were estimated from seven exposed cases aligning the occurrence of the ADE on the timeline ("*E*" represents the start points of exposure when a prescription is dispensed, "*O*" represents the occurrence of the ADE and the solid lines imply the induction time of the medicine). As the curve indicates, RR increases from when the patients start being exposed to the medicine and reaches the peak after certain days. Once the patient survives this time point, the RR begins to drop until the maximum induction period ends. A shorter time gap between the prescription and the ADE represents less exposure to the medicine and the following ADE

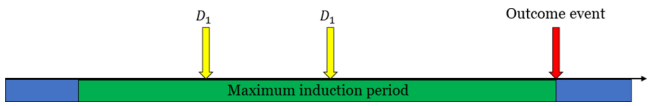

Fig. 5. Situation in which multiple prescriptions dispense the same medicine and the outcome event occurs.

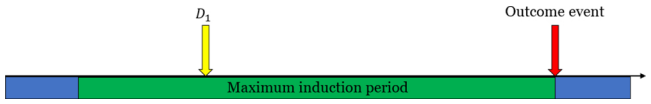

Fig. 6. Situation in which the time gap between the prescription and the outcome event is longer than that in the common situation.

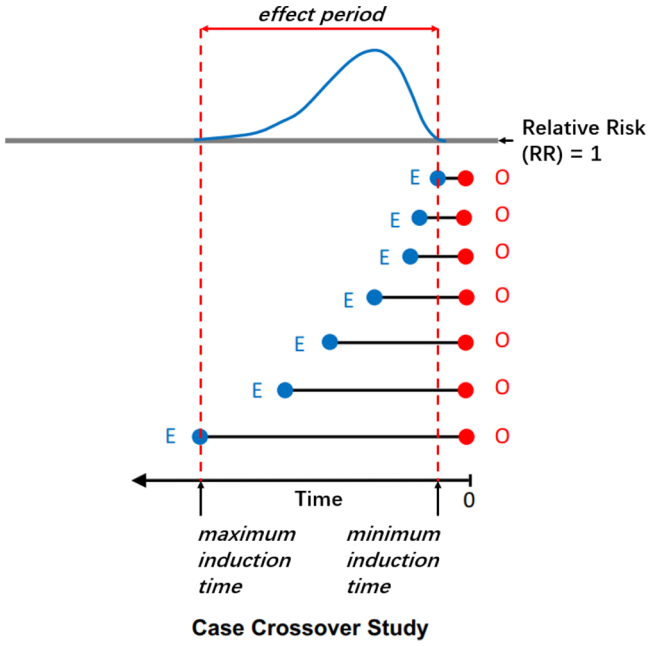

Fig. 7. Curve of relative risk regarding the occurrence of ADE in the case-crossover study.

may be caused by other reasons. A time gap of an appropriate length indicates that the influence of the medicine has accumulated to a certain level, thereby causing the ADE. In contrast, a prolonged time gap implies the patient might have already got rid of the influence of the medicine, and hence the association between this medicine and the ADE becomes questionable. In summary, the time gap between the prescription and the ADE reflects the suspicion of this medicine use on the ADE. Therefore, such time gaps are crucial to the differentiation of ADE-associated situation of medicine utilisation from other confounding situations, and we refer to the time gap as the temporal effect of a prescription attributing to the occurrence of the ADE.

Clearly, these three examples (Figs. 4–6) are different. The temporal effect and the frequency of a prescription are related to the levels of exposure, which eventually affects the occurrence of the ADE. However, the basic solution uses binary values (0 or 1) to indicate whether a candidate medicine has been dispensed or not irrespective of the frequency of prescriptions and the time gap between the prescription and the ADE. Thus, it only enables the recognition of the existence of exposure by treating all the above-mentioned situations as the same rather than measuring the degree of exposure. This limits the accuracy of the basic solution in estimating the effects of the candidate medicines on the ADE and results in the a false discovery.

3.3.2. Weight function

Therefore, we propose a weight function to measure the degrees of exposure by taking the temporal effect and the frequency of the prescription into account; thus, a more accurate estimation of the effects of medicines on the ADE can be achieved.

The first step in creating the weight function for a candidate medicine is to learn how the RR of ADE fluctuates over different days after the prescription is dispensed. Therefore, for each medicine in the candidate set **D***, we conduct a statistical analysis of all the time gaps

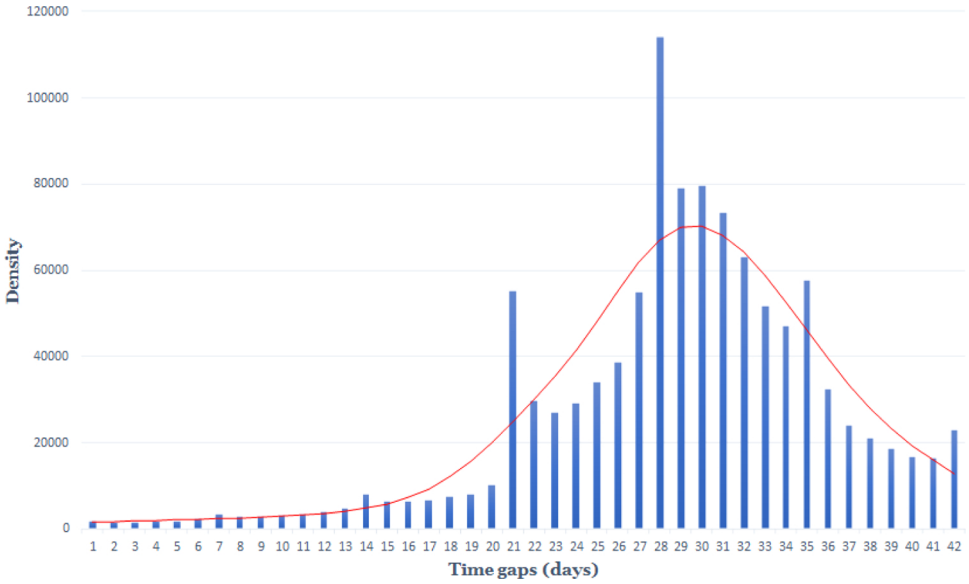

Fig. 8. Counts of time gaps between adjacent prescriptions that dispense vinblastine and aprepitant sequentially.

between the exposure (a prescription dispensing that medicine) and the outcome (the occurrence of the ADE) and calculate the counts of these time gaps (from day 1 to day τ in an ascending order). For example, Fig. 8 shows the counts of the time gaps between adjacent prescriptions that dispense vinblastine and aprepitant sequentially. Vinblastine is known for causing the adverse reactions of nausea and vomiting, which leads to the use of aprepitant for relief. Fig. 8 shows that the curve of the counts increase and drop, which is consistent with the curve of the previous research shown in 7 . The counts reach a peak on day 28, which implies that the maximum incidence of the ADE (nausea and vomiting) roughly occurs at day 28 after being exposed to vinblastine.

The purpose of the weight function is to facilitate the recognition of ADE-associated situation of medicine utilisation by rewarding correct patterns whose time gap between the antecedent medicine use and its succeeding ADE is of an appropriate length and penalising the false patterns with the time gaps of extreme lengths (too long or too short). More specifically, while the binary indicators ($\{0, 1\}$) in the basic solution can only distinguish whether the exposure (to a medicine) exists or not, the weight function measures the extents of exposure by converting binary indicators into continuous values between 0 and 1. To measure the varying RR of the ADE with different time gaps, i.e. the temporal effect, the weight function has to simulate the curve of the RR. Whereas being weighted as 1 represents the estimated maximum exposure to a medicine when the curve of RR of the ADE reaches its peak on a specific day, and the weights are scaled depending on the counts of the different time gaps.

Definition 4. Weight function: For medicine D_m and ADE $\mathcal{E}$, let $f_m(t)$ be the count of ADE $\mathcal{E}$ that occurs to patients on the tth day after receiving the prescription of D_m. The weight function $w_m(t)$ weights the temporal effect with respect to t, the length of time gaps between the prescription dispensing medicine D_m; and the end of the time window, and is defined as $w_m(t) = \frac{f_m(t)}{\max(f_m(t')) \mid t' \in [1,\tau]}$, where τ is the maximum length of the induction time regarding medicine use.

By substituting $w_m(t)$ for the medicine indicator d_m in both the case and the control time windows, the vector $\mathbf{V}_k = \{d_1, \ldots, d_m, \ldots, d_M\}$ is replaced with $\mathbf{W}_k = \{w_1(t), \ldots, w_m(t), \ldots, w_M(t)\}$. Therefore, the final optimisation problem of Eq. (3) is as follows:

$$\underset{\vec{\beta}}{\operatorname{argmin}} = -\frac{1}{K}\sum_{k=1}^{K}[y_k{}^*\mathbf{W}_k\vec{\beta} - \log(1 + e^{(\mathbf{W}_k\vec{\beta})})] + \lambda\|\vec{\beta}\|_1 \tag{4}$$

For the situation that there are multiple prescriptions dispensing the same medicine (D_m) in a time window, i.e. the frequency of prescription, we calculate the weight for each prescription and sum them all as the final value of the function, i.e.

$$\begin{aligned} w_m(t_1^i, \ldots, t_p^i, \ldots, t_P^i) &= \sum_{p=1}^{P} w_m(t_p^i) \\ &= \sum_{p=1}^{P} \frac{f_m(t_p^i)}{\max(f_m(t'))|t' \in [1,\tau]} \end{aligned} \tag{5}$$

where P is the number of prescriptions dispensing D_m in the time window and t_p^i is the length of the time gap between the pth prescription and the end of the time window for patient i.

Proof. Suppose that there are P sampled prescriptions in a time window denoted as $d_1, \ldots, d_P$, respectively; then, by using the weight function we can calculate weights for each record, i.e. $w_1, \ldots, w_P$. According to the setting of the logistic regression model, we assign the coefficients β_p to each weight of the prescription to quantify their effects on the ADE of the outcome; hence, the z in Equation (1) can be expressed as follows:

$$\begin{aligned} z(W) &= \beta_1 w_1 + \cdots + \beta_p w_p + \cdots + \beta_P w_P + \epsilon \\ &= \sum_{p=1}^{P} \beta_p w_p + \epsilon, \\ &\epsilon \sim N(0, \sigma^2) \end{aligned} \tag{6}$$

Presuming the effect of the same medicine on the ADE of the outcome is fixed (denoted as β'), when the above-mentioned prescriptions dispense the same medicine, we obtain $\beta_1 = \beta_2 = \cdots = \beta_P = \beta'$. Therefore, z can be written as $z(W) = \beta' \sum_{p=1}^{P} w_p + \epsilon$, which is the coefficient β' times the sum of all the weights regarding each prescription. In conclusion, the weight function for multiple prescriptions dispensing the same medicine is equal to the sum of the weight functions for all the individual prescription. □

We use the situation in Fig. 5 as an example to illustrate the process of applying the weight function. Suppose that the time gap between the first prescription dispensing D_1 and the occurrence of the outcome even is 35 days, and the gap between the second prescription dispensing D_1 and the event is 28 days. According to the statistic of delays of the ADE occurring to the patients after the dispensation of D_1 from historical data, the count of patients encountering the ADEs on day 28 reaches the peak, and then, drops to the half of the peak on day 35. Therefore, the second prescription should be weighted as 1 and we weight the first

prescription as 0.5. The final value of the weight function in terms of this example is the sum of weights of these two prescriptions, i.e. $1 + 0.5 = 1.5$.

In the meantime, medicines tend to be dispensed successively and comply with different resupply cycles. This results in medicines with shorter resupply cycles to have more prescriptions in the same time window. To adjust such an imbalance and avoid a possible bias towards the infrequently used medicines, we normalise the final values of the weight function into the same range (i.e. [0, 1]) by medicines. More specifically, for each candidate medicine d_m and final values $w_m(t_p^i)$ that the weight function calculated for patient $i \in \{1, ...,N\}$, we map $w_m(t_p^i)$ from $[\min(w_m(t_p^i)), \max(w_m(t_p^i))]$ to [0, 1] by scale.

3.4. Considering individual effect on ADE

In the case-crossover study, by each patient serving as his/her own control, we actually conduct a self-matching on each patient. In other words, a patient in the analysis is a stratum that consists of sampled medicine use information with both positive and negative outcomes. In the basic solution, the model does not distinguish medicine use information from different patients/strata. However, patients might respond differently to the same medicine, resulting from the varying characteristics among the patients. Thus, the patients have different baseline probabilities of encountering the corresponding ADE. To model such difference between patients, we propose to consider the individual effects that contribute to the occurrence of ADE by adding an additional parameter α to the basic model. Hence, after the addition of the individual parameter, the probability of Equation (1) can be expressed as follows:

$$\Pr(y_i^j|\mathbf{V}_i^j) = \frac{1}{1 + e^{-(\alpha_i + \mathbf{V}_i^j\vec{\beta} + \epsilon)}} \tag{7}$$

The additional parameter α_i can be interpreted as the reference log odds of patient i upon encountering the ADE ($y = 1$) versus not encountering it ($y = 0$) when all candidate medicines are absent ($d_m = 0|m \in \{1, ...,M\}$). As the individual parameters α_i vary across patients, to distinguish among different patients, we organise all the constructed time windows (both the case and the control time windows) by patients and use j to index the corresponding vectors of the medicine use information for each patient. Therefore, $\mathbf{V}_i^j$ represents the vector of the medicine use information that was sampled from the jth time windows of patient i. As we only use the initial prescription of dispensing the ADE-indicator medicine as the occurrence of ADE, for each patient, there is only one case time window and hence only one control time window are constructed, i.e. $j \in \{1, 2\}$.

Therefore, the final optimisation problem can be expressed as follows:

$$\begin{aligned}\text{argmin}_{\vec{\alpha},\vec{\beta}} \; = \; & -\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{2}[y_i^j*(\alpha_i + \mathbf{V}_i^j\vec{\beta}) \\ & - \log(1 + e^{(\alpha_i + \mathbf{V}_i^j\vec{\beta})})] + \lambda\|\vec{\beta}\|_1\end{aligned} \tag{8}$$

where $K = \sum_{i=1}^{N} 2 = 2N$ and y_i^j is the outcome of the corresponding time window ($y_i^j = 1$ if the j-th time window for patient i is a case time window and $y_i^j = 0$ when it is a control time window), which indicates whether an ADE occurs.

However, the parameter β is of our primary interest, while α is the nuisance parameter. To eliminate α from the equation, inspired by the work in [31], we let

$$\begin{aligned}\ell(\vec{\alpha},\vec{\beta}) \; = \; & -\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{2}[y_i^j*(\alpha_i + \mathbf{V}_i^j\vec{\beta}) \\ & - \log(1 + e^{(\alpha_i + \mathbf{V}_i^j\vec{\beta})})] + \lambda\|\vec{\beta}\|_1\end{aligned} \tag{9}$$

We consider that,

$$\frac{\partial\ell(\vec{\alpha},\vec{\beta})}{\partial\vec{\alpha}} = 0 \Rightarrow \alpha_i = \log\frac{\bar{y}_i}{1-\bar{y}_i} - \bar{\mathbf{V}}_i\vec{\beta} \tag{10}$$

where $\bar{y}_i = \frac{1}{2}\sum_{j=1}^{2} y_i^j \mid y_i^j \in \{0, 1\}$, which is the average outcome of all the case and the control time windows regarding patient i. And $\bar{\mathbf{V}}_i$ is a vector of M elements, whose each element is equal to the average value of the corresponding element from the vectors of the medicine use information $\mathbf{V}_i^j|j \in \{1, 2\}$ that belong to patient i.

As we balanced the number of case and control time windows (i.e. one case time window versus one control time window), the average outcome $\bar{y}_i = 0.5$, and thus, $\log\frac{\bar{y}_i}{1-\bar{y}_i} = 0$. Therefore, α_i can be written as follows:

$$\alpha_i = -\bar{\mathbf{V}}_i\vec{\beta} \tag{11}$$

By substituting Eq. (11) into (9), we can express the final optimisation problem which eliminated $\vec{\alpha}$ as follows:

$$\begin{aligned}\underset{\vec{\beta}}{\text{argmin}} \; = \; & -\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{2}[y_i^j*(\mathbf{V}_i^j - \bar{\mathbf{V}}_i)\vec{\beta} \\ & - \log(1 + e^{(\mathbf{V}_i^j - \bar{\mathbf{V}}_i)\vec{\beta}})] + \lambda\|\vec{\beta}\|_1\end{aligned} \tag{12}$$

Note that this improvement could be used together with the weight function. In such a case, $\bar{\mathbf{V}}_i$ will be replaced with $\bar{\mathbf{W}}_i$, which is a vector of M elements, each of which is equal to the average value of the corresponding element from the vectors of the weighted medicine use information $\mathbf{W}_i^j|j \in \{1, 2\}$, and the final optimisation problem can then be expressed as follows:

$$\begin{aligned}\underset{\vec{\beta}}{\text{argmin}} \; = \; & -\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{2}[y_i^j*(\mathbf{W}_i^j - \bar{\mathbf{W}}_i)\vec{\beta} \\ & - \log(1 + e^{(\mathbf{W}_i^j - \bar{\mathbf{W}}_i)\vec{\beta}})] + \lambda\|\vec{\beta}\|_1\end{aligned} \tag{13}$$

4. Experiments and results

The dataset utilised in the present paper is sourced from the claims of prescriptions under the Pharmaceutical Benefits Scheme (PBS) of Australia [32]. For each patient eligible for the PBS scheme, the PBS dataset records all the prescriptions of the patient without exception. All of the claims of the prescriptions in the PBS dataset are recorded using patient IDs and the dispensing date of the medicine, along with its generic name, Anatomical Therapeutic Chemical (ATC) classification code, and the PBS item code (an Australian code identifying the formulation). We applied the proposed method to a segment of the PBS dataset from 1 January 2013 to 31 December 2017.

4.1. Case study: signalling ADEs of heart failure and/or peripheral oedema

As the first part of the experiments, we conducted a case study by setting heart/cardiac failure and/or peripheral oedema as the target ADE. We choose the initial prescription dispensing frusemide (ATC code C03CA01) as the indicator of ADE. The use of furosemide is specific to the target ADE with less interference, and the previous study has

already shown that furosemide performs as a good proxy of heart failure [17]. In all, 48,984 patients in the dataset who had been dispensed with frusemide were included in the analysis.

To determine the length of τ (the estimated maximum exposure period of a medicine), we conducted a survey of the cycles of patients receiving the resupplied prescriptions by analysing the time gaps between the adjacent prescriptions dispensing the same medicine. A patient remained under the influence of a medicine as long as he/she kept taking it. Once the medicine ran out, the patient sought a prescription resupply if the medicine was still necessary. Therefore, this cycle of resupply indicates the length of the exposure period. According to our previous work, every seven days, there is a peak for patients getting the resupply, and it reaches the maximum frequency on day 28 (roughly 1 month). Within 42 days, a considerable proportion (90%) of the patients sought a resupply after the previous prescription. Therefore, for this case study, to include as many medicines into the candidate set as possible, we set the length of τ as 42 days. It is worth mentioning that we do not recommend to set τ as an extremely long length, e.g. 180 days. Such a setting violates the hypothesis of case time window as we can hardly blame the ADE to medicine that patients have taken a half year ago.

For locating the occurrences of the ADE by the initial prescription of frusemide, we constructed the case time windows accordingly, and in all, 804 medicines were assessed to be the candidate medicines. Relying on the criteria proposed in Section 3.1.3, we further selected a time window as the control time window for each patient. We sampled the medicine use information from both the case time windows and the control time windows for all the patients. Furthermore, as stated in Section 3.3, we calculated the counts of time gaps for each pair of a medicine in the candidate set and the indicator medicine from the entire dataset. Therefore, we customised the weight function w_m for each medicine in the candidate set and added them to the model. We also introduced the individual effects of the patients by adding the parameter α to the model. Eventually, we used these two improvements in combination and applied them together in the model. Then, we trained the logistic regression model with the Lasso regularisation from the sampled medicine use information. In the experiments of this work, the Lasso regression was implemented by *glmnet* [33], a sophisticated package on R. The *glmnet* package does not require users explicitly input the value of regularisation parameter lambda. It automates the process of regularisation parameter tuning by passing a grid of values for lambda and fits many models at once. When the model fitting and regularisation parameter tuning process ends, the *glmnet* package will also highlight the most well-trained and regularised model for users. Eventually, our method selected roughly 10% significant medicines from the candidate set and ranked them by coefficient β in the descending order as the output.

4.1.1. Results evaluation

Considering that it is impractical to find a benchmark database as "ground truth", we extracted the Australia approved product information (PI) from the Australian Therapeutic Goods Administration (TGA) website to review whether heart/cardiac failure and/or peripheral oedema were listed as an adverse event in the PI of the detected medicine. The information in a PI document was written by the pharmaceutical companies responsible for the medicine and has been approved by the TGA. It provides objective information about the quality, safety and effectiveness of the medicine, as demonstrated in the data provided to the TGA by the pharmaceutical companies [34]. If the PI explicitly listed heart/cardiac failure and/or peripheral oedema as the ADEs of a detected medicine, this medicine was considered a positive result. If the PI mentioned other heart/cardiac problems (other than heart/cardiac failure, such as heart/cardiac disorders) as an ADE of a detected medicine, this medicine was considered a neutral result. The medicines belonging to neutral results potentially suggest the new signals of the unknown relationships between the detected medicines and the heart failure, and our manual review shows that a great proportion of medicines in neutral results also have listed oedema (other than peripheral oedema) as an ADE in their product information. Besides, if the product information has not included any heart/cardiac problem or peripheral oedema (even other oedemas were mention), we have considered the corresponding medicine as a negative result.

In all, 86 signals of the medicine were detected, among which 48.84% (42/86) of the medicines were validated as positive results, 33.72% (29/86) were reviewed as neutral results, while only 17.44% (15/86) were negative results. We have listed the details of the detected signals of the positive results in Table 1.

4.1.2. Comparison with PSSA and SRS-based method

Currently, the PSSA is one of the most famous methods using prescription data for ADE detection. It can be considered as good baseline method. We compared the performance of our method with that of the PSSA via the same case study conducted by a by previous study [17]. In this study, authors have used PSSA as a signal detection tool in prescription claims data from Australia for detecting medicines with potential heart failure adverse event signals. They also used frusemide initiation as an indicator of heart failure. A signal was considered to be detected if the lower limit of the 95% confidence interval for the adjusted sequence ratio (the measurement of PSSA) was greater than one.

A total of 691 medicines were tested, and 22% (153/691) medicines were considered to be heart failure adverse event signals when using frusemide initiation as an indicator of heart failure. Their result evaluation was also conducted by referring product information provided by the Australian Therapeutic Goods Administration (TGA). Their statistics show that 60.78% (93/153) detected medicine whose product information listed heart failure or oedema as an ADE.

In summary, we have adapted stricter criteria for result evaluation in our work than the previous study did [17]. Approximate to the criterion used to verify the detected signals in [17], a total of 71 (42 + 29) medicines, in the prevalence of 82.56%, can be considered as valid ADE signals of heart/cardiac failure, peripheral oedema or other heart/cardiac problems. Therefore, our method outperformed the PSSA in the precision of detecting valid ADE signals of heart failure and/or peripheral oedema. Table 2 shows the comparison of the prevalence of valid signals detected by PSSA and our method.

To prove that our method complements the SRS, we have conducted an additional experiment to illustrate some ADE signals previously missed by SRS-based methods can be detected by our method.

We have implemented the proportional reporting ratios (PRRs) [5], a well-known SRS-based method, for the same case study, i.e. detecting medicines that might lead to heart/cardiac failure and/or oedema peripheral from FDA's SRS database. The SRS data is extracted from the FDA's database and we choose the exact time range as we did for evaluating our method, i.e. from 2013Q1 to 2017Q4. We also adapt the same minimum criteria for detecting a signal as used in the previous study [5], i.e. 3 or more cases, PRR at least 2 and 4 for the critical value

Table 1
Detected signals of positive results.

	ATC code	Generic name	Num in [a]	Num in controls[b]	Exclusive num in [c]
Proton pump inhibitors	A02BC02	Pantoprazole	6177	5978	1458
	A02BC05	Esomeprazole	6853	7075	1333
Antithrombotic agents	B01AB01	Heparin sodium	423	91	409
	B01AB05	Enoxaparin sodium	1702	855	1422
	B01AC16	Eptifibatide	3	0	3
	B01AC24	Ticagrelor	316	221	147
	B01AE07	Dabigatran	475	433	105
	B01AF01	Rivaroxaban	1634	1374	505
	B01AF02	Apixaban	1111	846	389
Antiarrhythmic medications	C01BD01	Amiodarone	925	620	436
Antihypertensives medications	C02AC05	Moxonidine	970	993	118
	C02KX04	Macitentan	15	10	7
Beta blocking agents	C07AB07	Bisoprolol	1282	982	437
	C07AB12	Nebivolol	287	202	121
Calcium channel blockers	C08CA01	Amlodipine	3542	3680	672
Antibacterials for systemic use	J01CR03	Ticarcillin and clavulanic acid	78	23	75
	J01DD04	Ceftriaxone	434	143	421
	J01FA06	Roxithromycin	1406	1368	1142
Antineoplastic agents	L01AA09	Bendamustine	4	1	3
	L01CB01	Etoposide	78	57	45
	L01CD01	Paclitaxel	228	186	86
	L01DB06	Idarubicin	5	0	5
	L01XC03	Trastuzumab	69	52	20
	L01XX32	Bortezomib	91	58	52
Analgesics	N02AA55	Oxycodone and naloxone	2041	1780	959
	N02AB03	Fentanyl	653	538	247
	N02AE01	Buprenorphine	1552	1407	464
Antiepileptics	N03AE01	Clonazepam	59	48	39
	N03AX16	Pregabalin	3371	3225	913
antipsychotic	N05AD01	Haloperidol	165	130	121
	N05AH03	Olanzapine	516	504	110
	N05AX08	Risperidone	656	631	221
Psychoanaleptics	N06AX11	Mirtazapine	1736	1718	389
Antifungal medications	J02AC04	Posaconazole	33	19	25
	A01AB04	Amphotericin B	254	224	218
Other medications	A12BA01	Potassium chloride	603	414	480
	B03XA02	Darbepoetin alfa	172	140	104
	C09DX03	Olmesartan, amlodipine and hydrochlorothiazide	136	128	29
	H02AB09	Hydrocortisone	138	65	108
	J05AB14	Valganciclovir	41	15	38
	R03AL04	Indacaterol and glycopyrronium bromide	165	142	61
	V03AF01	Mesna	27	15	17

[a] Number of patients who used the suspected medicine in their case time windows.
[b] Number of patients who used the suspected medicine in their control time windows.
[c] Number of patients who exclusively used the suspected medicine in their case time windows (not in their control time windows).

Table 2
Comparison between valid signals detected by the proposed method and the PSSA.

	PSSA	Proposed method	
Valid signals	60.78%	48.84% (Positive results) 33.72% (Neutral results)	82.56% in total
Invalid signals	39.22%	17.44% (Negative results)	

Table 3
Details of signal comparison detected by the proposed method and the SRS-based baseline method.

	Signals consistent with PPRs	Signals newly detected	Total	Prevalence (newly detected/total)
Positive results	29	13	42	30.95%
Neutral results	9	20	29	68.67%
Negative results	5	10	15	66.67%
Total	43	43	86	50%

of the chi-squared test. From this experiment, we have summarised the total of 620 signals detected by PPRs.

By comparing the signals detected by our method with the signals detected by PPRs, half (50%) of the signals detected by our method are overlapped with signals detected by PPRs, which indicates the reliability of our method. Besides, our method has detected 43 new signals (50% of signals detected by our method) which are not detected by PPRs. Among these 43 signals detected by our method only, 13 signals were evaluated as positive results, 20 signals were evaluated as neutral results and 10 signals were negative results. The details of the comparison of the signals are listed in Table 3.

4.2. Comprehensive ADE signal detection: gold standard-based experiments

As the second part of the experiments, to thoroughly assess the performance of the proposed method, we have conducted more comprehensive experiments according to a gold standard from a previous study [35]. According to the study, these 19 medicines were carefully selected by domain experts. The domain experts ensured these medicines were available for the Australian market consecutively between

Table 4
Sensitivity, specificity and predictive values of our method and improvements.

Associations	Gold standard		Total	
	Positive medicine	Negative medicine		
Positive prediction	27	2	29	Positive predictive value = 27/29 = 93.1%
Negative prediction	14	51	65	Negative predictive value = 51/65 = 78.5%
Total	41	53	94	
	Sensitivity = 27/41 = 65.9%	Specificity = 51/53 = 96.2%		

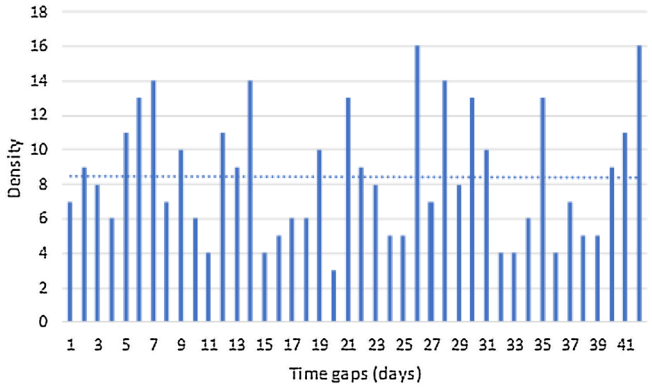

Fig. 9. Frequencies of time gaps between adjacent prescriptions that dispense ranitidine and tobramycin sequentially, which unrelated to each other. We randomly selected these two medicines as an example for the uniform distribution.

1995 and 2010. These medicines have a relatively high usage volume in the population, and this ensures an adequate sample size for research studies. Besides, the domain experts also ensured that these medicines were not predominantly used in a specific subpopulation. Eventually, this gold standard listed 19 medicines involving a great diversity in their anatomical, therapeutic and chemical properties. Benefited from such a selection, the gold standard contains 43 positive events which can be caused by these medicines and 114 negative events which are unlikely to be caused by these medicines. For example, perindopril indeed causes cough and dizziness, but considering seizure and colitis as ADEs of perindopril is not necessarily correct. From this gold standard, we have summarised a new standard containing 20 adverse events, along with their indicator medicines, 41 positive medicines which are capable of causing corresponding ADEs in the 20 adverse events and 53 negative medicines which are not. For example, cough, indicated by the use of pholcodine linctus, can be caused by perindopril and ramipril, while it is unlikely to be caused by strontium ranelate. Comparing to the previous case study, these comprehensive experiments cover a broader range of ADEs, including common ADEs such as nausea and rare ADEs such as seizure.

The same PBS prescription dataset was utilised for these experiments. Our method successfully detected 65.9% (27/41) positive medicines, while only 3.8% (2/53) negative controls were mistakenly detected (i.e. false positives). The sensitivity, specificity, positive predictive value and negative predictive value are listed in Table 4, and more details along with the gold standard are provided in Table 5 in Appendix A.

4.2.1. Comparison to state-of-the-art methods

We compared the performance of the proposed method with two previous studies using the same prescription dataset and conducting the evaluation according to the same gold standard. The first study [36] is the validation of effectiveness of the PSSA on Australia Prescription data (PBS data). The PSSA achieved a sensitivity of 61% and a specificity of 93%. Another study [37] is a supervised machine learning methods for detecting ADEs from prescription data, published by 2018. They spilt the standard for training and testing, and their results ranged from 56% to 84% (average 67.8%) in sensitivity and 50% to 90% (average 74%) in specificity, depending on different machine learning techniques (such as logistic regression, decision tree, SVM, random forest, Neural network). When they used another standard for training and the above-mentioned standard for testing, the results ranged from 48% to 70% (average 50.8%) in sensitivity and 20% to 89% (average 65.7%) in specificity. In summary, with a sensitivity of 65.9% and specificity of 96.2%, the proposed method has achieved a promising performance comparing with existing methods.

5. Discussion

In this section, we plan to discuss some interesting things that are related to our method, experiment results and other existing researches.

First of all, we would like to discuss the confounding problem related to this study. First, as the PBS dataset provides a complete chain of prescription for each patient without missing records, we are less concerned about unobserved confounders raised from additional medicines. Second, the use of case-crossover study as data preparation for logistic regression analysis eliminates the confounding situations raised from time-invariant characteristics of patients. The criterion of “sufficient closeness” for the selection of control time windows also reduces the influence of confounding situations raised from time-variant patient-specific confounders. Third, although the dispensation of an ADE-indicator medicine could be caused by a new indication/disease other than the development of an ADE, this would not cause a confounding situation in utilising prescription data. Despite that the association between the new disease and the ADE-indicator medicine is certain, the dispensation of investigated medicine is independent of the new disease. Therefore, a new disease would not influence the detection of ADE-causing medicine in prescription data because it would not generate any spurious associations between the investigated medicine and the ADE-indicator medicine. Fourth, the co-exposures [38] problem may introduce a confounding situation to our study, i.e. one medicine is frequently co-prescribed with another medicine, if any of them causes an ADE then its effect on the ADE may be masked by the other. Such a situation might lead to false discovery (the signal of co-prescribed medicine), it never prevents the true signal of ADE from being detected by our method. Fifth, there may exist the confounding situation raised from acute indication [38], i.e. an earlier disease/symptom leads to the dispensation of a medicine as the treatment and after that, the earlier disease/symptom develops into another disease/symptom which can be mistakenly considered as the ADE of the

treatment medicine. The weight function we proposed in Section 3.3 takes the time gaps between the prescription and the ADE into account. By rewarding the sampled medicine use information whose time gap between the prescription and the ADE complies with the temporal pattern of ADE-associated situation and penalising the sampled medicine use information whose time gap does not comply, our method distinguishes the ADE-associated situation from other confounding situations and mitigates the influences of such confounders.

In order to detect more potential ADE signals, we use the estimated maximum exposure period as the length of sampled (case and control) time windows as we proposed in Section 3. In the case study that setting heart/cardiac failure and/or peripheral oedema as target ADE, potential signals we detected cover a variety of medicines, including antithrombotic agents (ATC code initialled by B01A), antineoplastic agents (ATC code initialled by L01) and analgesics (ATC code initialled by N02), etc. (details listed in Table 1) were detected. In practice, a uniform length of sampled time windows cannot fit all medicines. While a longer time window ensures the higher sensitivity, the compromise of specificity comes along. The false discoveries of our case study generally focus on antibiotics (ATC code initialled by J01), which are usually intermittently used by patients. Hence, such signals were captured as those medicines are less prone to be sampled in a control time window (patients are in a healthy condition). Besides, any settings of time window length shorter than 42 days are also welcomed if users aim to detect more acute events caused by medicine use with shorter latency. We have conducted a few more experiments of different length of time windows, i.e. 42 days, 28 days and 14 days respectively. The results show that different length of time windows (28 days and 14 days) facilitate the detection of more signals which were not captured with the maximum length of the exposure period (42 days), such as blood glucose lowering medications (ATC code initialled by A10B) and immunosuppressants (ATC code initialled by L04A), etc. However, a narrow time window also brings additional false signals, e.g. some medicines are usually co-prescribed with indicator medicine (frusemide in our case study).

In the experiments of comparing to the gold standard, the two improvements we proposed with our basic method have shown their effectiveness in a more comprehensive evaluation. The sensitivity and specificity of results were gradually enhanced with the addition of two improvements. The performance of the method overwhelms that of the basic solution when two improvements are used in combination. Speaking to adding the individual effects, it automatically adjusts the corresponding indicators of medicine use information to 0 when a candidate medicine was evenly dispensed to a patient during both his case and control time windows. Thus, unrelated medicines were eliminated as those medicines are considered to be less possible to lead to the ADE. There are also a number of positive medicines in the gold standard which our method failed to detect. Two major reasons might be responsible for those false negatives: (1) Signals of some medicines are masked by medicines belonging to the same pharmaceutical class but are more frequently used. Frequently used medicines provide more cases that patients experiencing the ADE, and hence generate stronger signals. When the method selected strong signals among all the candidates, frequently used medicines tend to mask other medicines with lower usages, especially those medicines belonging to the same pharmaceutical class. For example, we successfully detected Sertraline as causing the ADE, nausea. However, we failed to detect Citalopram and Escitalopram as the cause of nausea, which belong to the same pharmaceutical class (selective serotonin reuptake inhibitors, ATC code prefix N06AB) but are less frequently used. (2) Some medicines are not sufficiently sampled in case time windows due to the uniform setting of the length of time windows in the experiments. As stated in the previous paragraph, the uniform setting of the length of time windows may not fit for all medicines. It requires more evaluations to determine the optimal individualised settings.

It is also worth to mention a phenomenon we discovered when surveying the time gaps between the dispensations of medicines for the weight function. That is, for two medicines that one causes the ADEs which lead to the dispensation of another, the counts regarding different values of their time gaps tend to follow the curve as Fig. 8 shows. However, for two medicines that do not have a strong association, the counts will roughly comply with a uniform distribution as Fig. 9 shows. This phenomenon reminds us of the PSSA [15] method, which evaluates the association between suspected medicines and the ADE-indicator medicine by measuring the asymmetry of suspected medicines being dispensed before versus after the dispensation of ADE-indicator.

Last but not least, the main value of our method lies in its high efficiency in detecting the most suspicious signals of ADEs from the large-scale prescription data. So the patients and medical practitioners can be alerted of possible ADEs and possible consequences of ADEs can be prevented. Despite that our experiments were conducted on a prescription dataset which does not contain additional information such as symptoms, diagnoses nor physical measurements, our method has the potential to be a complementary tool for global pharmacosurveillance relying on SRS. For future work, we plan to extend our method to automatically utilise domain knowledge to improve the process of ADE detection.

6. Conclusion

In this paper, we introduced a method to discover medicines that are responsible for a specific ADE from the prescription data by spontaneously evaluating their effects on the corresponding ADE. For data preparation, we adapt the design of the case-crossover study to construct the case and the control time windows for the extraction of medicine use information. Furthermore, we proposed several improvements to take into account more factors (i.e. the temporal effect, frequency of prescription and individual effect) that might affect the evaluation. Therefore, some biases were eliminated and a more accurate discovery was achieved.

Conflict of interest

None declared.

Acknowledgements

This research was funded by the Australian Research Council Australia (ARC)DP170101306 and the National Health and Medical Research Council Australia (NHMRC)GNT1040938. We acknowledge student support through the Australian Government Research Training Program Scholarship and RTIS seed funding of the University of South Australia and ER is supported by NHMRC GNT1110139.

Appendix A. Appendix

Table 5 provides the gold standard we summarised and marks the detail of the experimental results of comparing to the gold standard.

Table 5
Summarised gold standard and details of our experimental results.

ADE	Indicator Medicine	ATC of Indicator Medicine	Positive Medicine	ATC of Positive Medicine	Negative Medicine	ATC of Negative Medicine
Nausea	Metoclopramide, domperidone	A03FA01, A03FA03	Metformin	A10BA02		
			Strontium ranelate	M05BX03		
			Donepezil	N06DA02		
			Tramadol	N02AX02		
			Venlafaxine	N06AX16		
			Citalopram	N06AB04		
			Escitalopram	N06AB10		
			Sertraline	N06AB06		
						None
Oedema/peripheral oedema	Frusemide	C03CA01	Amlodipine	C08CA01		
			Risperidone	N05AX08		
			Raloxifene	G03XC01		
			Celecoxib	M01AH01 / L01XX33		
						None
Diarrhoea	Loperamide hydrochloride, diphenoxylate hydrochloride	A07DA03, A07DA01	Carvedilol	C07AG02		
			Metformin	A10BA02		
			Rabeprazole	A02BC04		
			Donepezil	N06DA02		
			Raloxifene	G03XC01		
			Meloxicam	M01AC06		
			Celecoxib	M01AH01 / L01XX33		
			Escitalopram	N06AB10		
			Sertraline	N06AB06		
						None
Constipation	Laxatives	A06A	Tramadol	N02AX02		
			Venlafaxine	N06AX16		
			Mirtazapine	N06AX11		
					Raloxifene	G03XC01
Dyspepsia	Proton pump inhibitors	A02BC	Meloxicam	M01AC06		
			Celecoxib	M01AH01 / L01XX33		
			Sertraline	N06AB06		
Cough	Pholcodeine linctus	R05DA08	Perindopril	C09AA04		
			Ramipril	C09AA05	Strontium ranelate	M05BX03
Dizziness	Prochlorperazine	N05AB04	Perindopril	C09AA04		
			Irbesartan	C09CA04		
			Carvedilol	C07AG02		
			Risperidone	N05AX08		
			Donepezil	N06DA02		
			Tramadol	N02AX02		
			Venlafaxine	N06AX16		
			Mirtazapine	N06AX11		
			Escitalopram	N06AB10		
						None
Insomnia	Flunitrazepam, nitrazepam, temazepam, zopiclone	N05CD03, N05CD02, N05CD07, N05CF01	Donepezil	N06DA02		
			Tramadol	N02AX02	Metformin	A10BA02
Extrapyramidal symptoms	Benzhexol, biperiden, benztropine	N04AA01, N04AA02, N04AC01	Risperidone	N05AX08	Carvedilol	C07AG02
					Strontium ranelate	M05BX03
Seizure	Phenytoin	N03AB02			Perindopril	C09AA04
					Ramipril	C09AA05
					Irbesartan	C09CA04
					Carvedilol	C07AG02
				None	Raloxifene	G03XC01
Osteoporosis	Bisphosphonate, Strontium ranelate	M05BA, M05BB, M05BX			Perindopril	C09AA04
					Ramipril	C09AA05
					Irbesartan	C09CA04
					Amlodipine	C08CA01
					Carvedilol	C07AG02
					Metformin	A10BA02
				None	Donepezil	N06DA02
Colitis	Mesalazine	A07EC02			Perindopril	C09AA04
					Ramipril	C09AA05
					Irbesartan	C09CA04
					Amlodipine	C08CA01
					Metformin	A10BA02
					Strontium ranelate	M05BX03
					Risperidone	N05AX08
					Donepezil	N06DA02
					Raloxifene	G03XC01
				None	Tramadol	N02AX02
Migraine	Ergot alkaloids, selective 5HT receptor agonist	N02C			Metformin	A10BA02
					Strontium ranelate	M05BX03
					Risperidone	N05AX08
					Donepezil	N06DA02
					Meloxicam	M01AC06
				None	Celecoxib	M01AH01 / L01XX33
Depression	Antidepressions	N06A		None	Strontium ranelate	M05BX03
Hyperlipidemia	Statins	C10AA			Strontium ranelate	M05BX03
					Meloxicam	M01AC06
				None	Celecoxib	M01AH01 / L01XX33
Hyperuricemia/ gout	Colchicine	M04AC01			Strontium ranelate	M05BX03
					Donepezil	N06DA02
					Raloxifene	G03XC01
				None	Tramadol	N02AX02
Glaucoma	Antiglaucoma	S01E			Strontium ranelate	M05BX03
					Risperidone	N05AX08
					Meloxicam	M01AC06
				None	Celecoxib	M01AH01 / L01XX33
Urinary incontinence	Oxybutynin	G04BD04			Strontium ranelate	M05BX03
					Rabeprazole	A02BC04
				None	Raloxifene	G03XC01
Hyperglycemia	Insulin	A10A			Strontium ranelate	M05BX03
					Raloxifene	G03XC01
				None	Tramadol	N02AX02
Asthma/ bronchospams	Salbutamol	R03AC02, R03CC02			Raloxifene	G03XC01
				None	Tramadol	N02AX02

References

[1] Stausberg J, Hasford J. Identification of adverse drug events: the use of icd-10 coded diagnoses in routine hospital data. Deut Arztebl Int 2010;107(3):23.
[2] Johnson J, Bootman J. Drug-related morbidity and mortality: a cost of illness model. Arch Intern Med 1995;155:1949–56.
[3] Ahmad SR. Adverse drug event monitoring at the food and drug administration. J Gen Intern Med 2003;18(1):57–60.
[4] van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002;11(1):3–10.
[5] Evans S, Waller PC, Davis S. Use of proportional reporting ratios (PRRS) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001;10(6):483–6.
[6] Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 2004;13(8):519–23.
[7] Bate A. Bayesian confidence propagation neural network. Drug Saf 2007;30(7):623–5.
[8] Madigan D, Simpson S, Hua W, Paredes A, Fireman B, Maclure M. The self-controlled case series: recent developments. Columbia University; 2015.
[9] Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf 2009;18(12):1176–84.
[10] Wahab I, Pratt N, Kalisch L, Roughead E. Sequence symmetry analysis and disproportionality analyses: what percentage of adverse drug reaction do they signal.

Adv Pharmacoepidemiol Drug Saf 2013;2(140):1052–2167.
[11] Katz N, Panas L, Kim M, Audet AD, Bilansky A, Eadie J, et al. Usefulness of prescription monitoring programs for surveillance-analysis of schedule ii opioid prescription data in massachusetts, 1996–2006. Pharmacoepidemiol Drug Saf 2010;19(2):115–23.
[12] Crystal S, Akincigil A, Bilder S, Walkup JT. Studying prescription drug use and outcomes with medicaid claims data strengths, limitations, and strategies. Med Care 2007;45(10 SUPL):S58.
[13] Zhan C, Roughead E, Liu L, Pratt N, Li J. A data-driven method to detect adverse drug events from prescription data. J Biomed Informatics 2018;85:10–20.
[14] Petri H, De Vet H, Naus J, Urquhart J. Prescription sequence analysis: a new and fast method for assessing certain adverse reactions of prescription drugs in large populations. Stat Med 1988;7(11):1171–5.
[15] Hallas J. Evidence of depression provoked by cardiovascular medication: a prescription sequence symmetry analysis. Epidemiology 1996:478–84.
[16] Hersom K, Neary MP, Levaux HP, Klaskala W, Strauss JS. Isotretinoin and antidepressant pharmacotherapy: a prescription sequence symmetry analysis. J Am Acad Dermatol 2003;49(3):424–32.
[17] Wahab IA, Pratt NL, Ellett LK, Roughead EE. Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database. Drug Saf 2016;39(4):347–54.
[18] van Boven JF, Vegter S. Inhaled corticosteroids and the occurrence of oral candidiasis: a prescription sequence symmetry analysis. Drug Saf 2013;36(4):231–6.
[19] Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 1991;133(2):144–53.
[20] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B Methodol 1996;58(1):267–88.
[21] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342(25):1887–92.
[22] Schlesselman JJ. Case-control studies: design, conduct, analysis. Oxford University Press; 1982.
[23] Maclure M, Mittleman MA. Should we use a case-crossover design? Annu Rev Public Health 2000;21(1):193–221.
[24] Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health 2005;95(S1):S144–50.
[25] Kuhn M, Letunic I, Jensen LJ, Bork P. The sider database of drugs and side effects. Nucleic Acids Res 2015;44(D1):D1075–9.
[26] Granger CW. Causality, cointegration, and control. J Econ Dyn Control 1988;12(2-3):551–9.
[27] Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Methodol 1958:215–42.
[28] Pearl J. Causal inference in statistics: an overview. Stat Surv 2009;3:96–146.
[29] Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. Epidemiology 2017;28(2):237–48.
[30] Maclure M. Using simulated data to assess case-crossover designs for studying less transient effects of drugs. Drug Saf 2017:1–4.
[31] Xu S, Zeng C, Newcomer S, Nelson J, Glanz J. Use of fixed effects models to analyze self-controlled case series data in vaccine safety studies. J Biometrics Biostat 2012:006.
[32] Australian Government Department of Health, Pharmaceutical Benefits Scheme and Repatriation Pharmaceutical Benefits Scheme Section 85 Supply Data. http://www.pbs.gov.au.
[33] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1–22http://www.jstatsoft.org/v33/i01/.
[34] A. T. G. A. (TGA). Product information in TGA website. 2019https://www.tga.gov.au/product-information-0.
[35] Wahab IA, Pratt NL, Wiese MD, Kalisch LM, Roughead EE. The validity of sequence symmetry analysis (SSA) for adverse drug reaction signal detection. Pharmacoepidemiol Drug Saf 2013;22(5):496–502.
[36] Pratt NL, Ilomäki J, Raymond C, Roughead EE. The performance of sequence symmetry analysis as a tool for post-market surveillance of newly marketed medicines: a simulation study. BMC Med Res Methodol 2014;14(1):66.
[37] Hoang T, Liu J, Roughead E, Pratt N, Li J. Supervised signal detection for adverse drug reactions in medication dispensing data. Comput Methods Programs Biomed 2018;161:25–38.
[38] Mittleman MA, Mostofsky E. Exchangeability in the case-crossover design. Int J Epidemiol 2014;43(5):1645–55.