Detecting Adverse Drug Reactions in the General Practice Healthcare Database

Jenna Marie Reps

School of Computer Science

University of Nottingham

A thesis submitted for the degree of

Doctor of Philosophy

2014

Acknowledgements

I would like to acknowledge and thank my supervisors Professor Jonathan Garibaldi and Professor Uwe Aickelin, whos guidance made this thesis possible, my family for their continued support during my PhD and my friends for helping to keep me motivated.

Abstract

The novel contribution of this research is the development of a supervised algorithm that extracts relevant attributes from The Health Improvement Network database to detect prescription side effects. Prescription drug side effects are a common cause of morbidity throughout the world. Methods that aim to detect side effects have historically been limited due to the data available, but some of these limitations may be overcome by incorporating longitudinal observational databases into pharmacovigilance. Existing side effect detecting methods using longitudinal observational databases have shown promise at becoming a fundamental component of post marketing surveillance but unfortunately have high false positive rates. An extra step is required to further analyse and filter the potential side effects detected by existing methods due to their high false positive rates, and this reduces their efficiency. In this thesis a novel methodology, the supervised adverse drug reaction predictor (SAP) framework, is presented that learns from known side effects, and identifies patterns that can be utilised to detect unknown side effects. The Bradford-Hill causality considerations are used to derive suitable attributes as inputs into a learning algorithm. Both supervised and semi-supervised techniques are investigated due to the limited number of definitively

known side effects. The results showed that the SAP framework implementing a random forest classifier outperformed the existing methods on The Health Improvement Network longitudinal observational database, with AUCs ranging between 0.812-0.937, an overall MAP of 0.667, precision values between 0.733-1 and a false positive rate $\leq$ 0.013. When applied to the standard reference the SAP framework implementing a support vector machine obtained a MAP score of 0.490, an average AUC of 0.703 and a false positive rate of 0.16. The false positive rate is lower than that obtained by existing methods on the standard reference.

Contents

List of Figures

Nomenclature

(α, β)	denotes a drug-medical event pair, where the drug is α and the medical event is β.
ADE	Adverse drug event.
ADR	Adverse drug reaction.
AUC	Area under the ROC curve: a measure of general signalling ability.
DOI	Drug of interest.
HOI	Health outcome of interest.
HUNT	Highlighting UTARs Negating TARs: a method for signalling ADRs using LODs.
IC	Information component: the measure of association used by the TPD method.
$\mathbf{IC}_{\Delta}$	A measure of association change over time used by the TPD method.
KNN	K-nearest neighbour.

LOD Longitudinal observational database.

LR Logistic regression.

MAP Mean average precision.

MUTARA Mining Unexpected Temporal Association Rules given the Antecedent: a method for signalling ADRs using LODs.

NB Naive Bayes.

OMOP Observational Medical Outcomes Partnership.

pAUC Area under the partial ROC curve: a measure of signalling ability for a defined specificity interval.

RF Random forest.

RME Risk medical events: the set of medical event that are observed for at least one patient within a month of being prescribed the drug of interest.

ROC Receiver operating characteristic: an illustration of the performance of a binary classifier.

ROR Reporting odds ratio: a measure of association.

SAP Supervised ADR predictor: the novel framework for signalling ADRs developed through this thesis.

SIDER A side effect resource containing drug-medical event pairs corresponding to ADRs.

SRS Spontaneous reporting system.

SSAP Semi-supervised ADR predictor: a novel framework for signalling ADRs developed throughout this thesis.

SVM Support vector machine.

THIN The Health Improvement Network.

TPD Temporal Pattern Detection: an ADR signalling method developed for LODs that looks for temporal changes in the association strength between and drug and medical event.

WHO World Health Organization.

Chapter 1

Introduction

The occurrence of negative side effects due to prescribed medication is a health issue that occurs worldwide. The early detection of side effects is imperative for the prevention of unnecessary morbidities or mortalities. Two types of electronic healthcare databases are frequently used to extract data for the detection of side effects, the spontaneous reporting system (SRS) databases and the longitudinal observational databases (LODs). Many methods have been developed for the SRS databases but these databases have a limited perspective and do not contain the data required to detect all side effects. This has prompted the focus towards using the LODs, but the proposed methods tend to be unsupervised. In this thesis, supervised and semi-supervised techniques capable of detecting side effects by utilising the data contained in LODs are investigated. The first part of this chapter focuses on the research background and motivation, this is followed by the aims and objectives. The chapter concludes with the thesis organisation that provides the outline of each chapter.

1.1 Background & Motivation

All prescribed drugs have side effects under certain conditions [170]. A negative effect following the ingestion of a drug is referred to as an Adverse Drug Event (ADE) and is defined as 'any untoward medical occurrence that may present during treatment with a medicine but which does not necessarily have a causal relationship with the treatment' [206]. When an ADE has been linked to a specific drug it becomes an Adverse Drug Reaction (ADR). An ADR is defined by the World Health Organization as a response to a medicine which is noxious and unintended, and which occurs at doses normally used in humans from the prophylaxis, diagnosis or therapy of disease, or for the modification of physiological function [182].

As ADRs can lead to patient morbidity or mortality, their early discovery is essential. As a consequence, the safety of a new drug is extensively analysed throughout its development. Unfortunately, the ability to analyse a drug's toxicity is limited by the clinical study designs. The pre-clinical studies of a drug's development, involving animal testing, are done to initially assess a drug's toxicity [25], however, the ability to infer ADRs is limited by the inability of animal testing to be completely informative for effects on humans [133]. If a drug passes the initial toxicity analysis, it is then tested on humans during phases i–iii, with the trial population size increasing incrementally after each phase. Phase i will often involve testing the drug by giving it to healthy individuals under unrealistic conditions (i.e., the individuals cannot smoke, drink alcohol, exercise excessively and may have food limitations enforced) [38]. It is also widely known that clinical trials can be biased towards certain demographics, for example the majority

of individuals tested during phase i trails are white males [38]. Clinical trials involve testing the drug on a limited population size, with the largest population size used during phase iii, but this generally only contains up to 3000 individuals [38]. Due to numerous reasons, including the limited trial population size and the unrealistic conditions of the trials, many ADRs are undetectable during phase i-iii studies and can only be identified after the drug is marketed [11]. It is also clear that ADRs that result from polypharmacy (i.e., when multiple drugs are prescribed at the same time) will be difficult to detect. The reason is, due to the limited population being tested, it is impossible to investigate all the different drug combinations.

Studies investigating the prevalence of ADRs have provided evidence that many ADRs are not discovered prior to marketing. The results indicate that up to 6.5% of UK hospital admissions are due to ADRs [135] with similar rates also being observed in the US hospitals (6.7%) [103]. Another study found a similar prevalence within a UK paediatric hospital (4%) [62]. Research suggests ADRs are more common in geriatric patients (older than sixty five), in females and in patients taking more than one drug [17]. It has also been highlighted that the lack of efficient means to detect ADRs causes a burden in terms of cost and quality of life. Furthermore, this burden appears to be getting worse. It has been reported that ADRs could cost the UK £637 million each year [43], with £466 million being due to ADR hospital admissions and £171 million being due to ADRs during hospitalisation. These estimates do not take into consideration additional medical costs or loss of earning while in hospital due to an ADR. A study by Wu *et al.* (2010) compared the frequency of ADRs as the cause of hospital admission over 1999 to 2009 and showed the number of people admitted to hospital due to

ADRs has increased over the ten years at a greater rate than the rate of hospital admission [207]. Further, they found 26,399 people died in hospital in the UK over the ten years as a result of an ADR [207]. This corresponded to a probability of almost one in twenty ADRs resulting in death. One explanation for the increase in the number of ADRs over the years is due to polypharmacy [112].

This highlights the importance of continuous post-marketing surveillance of drugs and motivates the requirement of new methods that can identifying ADRs efficiently. When a new potential ADR association is detected, the potential ADR is referred to as being signalled. The majority of current post-marketing surveillance techniques make use of the SRS databases. These databases contain records of suspected ADRs, that were originally restricted to submissions made by medical practitioners and coroners, but it is becoming increasingly common for them to enable the general public to submit reports. The SRS databases have many limitations that prevent them signalling ADRs efficiently and they cannot be used to quantify ADR risks [57], nor can they be used to consistently identify risk factors. It is widely known that the majority of ADRs signalled by the existing methods applied to SRS databases do not correspond to ADRs [172]. Retrospective studies have confirmed their inability to efficiently signal all ADRs, as the methods applied to SRS databases were unable to signal some ADRs before they were discovered by other means [3]. This has prompted the demand for better surveillance techniques [43] [207] and to use other forms of data to complement drug safety using SRS.

An alternative approach for signalling ADRs, that has recently surfaced, is to use data contained within LODs. The LODs are not restricted to a specific period of time around the drug prescription and can contain patient medical histories

spanning decades. These databases may present to opportunity to efficiently discover new ADRs [198] and enable ADR risks to be quantified. Their importance for future post-marking surveillance has been expressed [203]. The existing methods proposed for the LODs are unsupervised and many are derived from the SRS methods [216] but new methods have been presented that are based on epidemiology techniques [156]. Unfortunately, these methods have been shown to have a high false positive rate [156], due to the difficulty distinguishing between association and causation, and this may reduce their signalling efficiency. The majority have been developed for a common data model [115; 131] (the integration of multiple LODs into a general database) rather than specific databases. Not all data can be converted into the common model [214], so information may be lost. Therefore, it is of interest to also develop methods that are specific to a single database, as new information may be revealed. It may be possible to develop a method with a low false positive rate by considering the Bradford-Hill causality consideration [19], as these are often used within extensive post marketing investigations to confirm causality .

The Health Improvement Network (THIN) database is an example of a LOD that contains approximately 6% of the general practice records within the UK. The THIN database contains complete medical records and prescription records (while the patient is registered) for all registered patients at participating practices. The THIN database contains heterogeneous data and has hierarchal structures embedded within it. An example of one of the hierarchal structures contained in the database is the recording of the medical events (i.e., administrative events, illnesses, symptoms, laboratory tests/results and medical history). The medical events are recorded via READ codes, these codes have five levels of

specificity and follow a tree structure. Little work to date has focused on using the THIN database for general postmarking surveillance and no ADR signalling method has been specifically developed. Research has suggested that the THIN database potentially holds a wealth of information [105]. If its complex structure can be dealt with, then its integration into post-marketing surveillance may enable ADRs to be signalled efficiently.

1.2 Aims & Objectives

As this research is interdisciplinary it has both a clinical and technical aim. The overall clinical aim of this project is to develop a data-mining algorithm for a specific LOD, the THIN database, that can detect ADRs and discover new information to improve current post marketing drug surveillance. The technical aim is to develop an algorithm that can classify a pair consisting of a drug and medical event (drug-medical event pair) as a causal relationship or non-causal relationship. The algorithm must have a low false positive rate and a sufficiently high true positive rate. This algorithm has multiple applications as it can identify causation in databases containing discrete information. One such example is using databases containing customer shopping histories to identify items that when purchased influence a different item being purchased in the future. Another useful implementation of the algorithm using market data could be to identify the impact of promotions and find what purchases are caused by the promotion. The advantage over sequential pattern mining is that the algorithm does not require the events to be common.

1.2.1 Hypotheses

The THIN database potentially contains a wealth of information but this is hidden within a magnitude of heterogeneous data containing many underlying hierarchal structures. The abilities of the existing ADR signalling methods developed for LODs are likely to be impacted by the structure of the THIN database and also by their inability to distinguish between association and causation. These limiting factors may prevent the extraction of all the information that is potential available by mining the THIN database. To extract all the possible information and utilise the full potential of the THIN database, novel supervised/semi-supervised methods may need to be developed. It is therefore hypothesised that,

H1 Current ADR signalling algorithms developed for LODs are not suitable for ADR detection when they are implemented on the THIN database.

H2 Novel ADR signalling algorithms applied to the THIN database will be able to consistently perform better than existing LOD ADR signalling algorithms if they 1) deal with the hierarchal structures within the THIN database, 2) incorporate new attributes essential for determining causality and 3) use known ADR knowledge.

H3 Novel ADR signalling algorithms applied to the THIN database will outcompete existing methods developed for the Observational Medical Outcomes Partnership (OMOP) common model when considering the specified drug and health outcomes of interest [141].

H4 Novel ADR signalling algorithms applied to the THIN database will be able to generate new ADR signals.

1.2.2 Objectives

To address the research hypotheses the following research objectives are proposed, with the hypothesis they are linked to indicated in brackets.

1. Determine the benchmark for signalling ADRs using the THIN database and identify limitations (**H1**).

2. Propose suitable attributes for each drug-medical event pair that may help separate association from causation or that are specific to the THIN database (**H2.1**-**H2.2**).

3. Develop a novel supervised/semi-supervised ADR signalling algorithm for implementation on the THIN database that can accurately signal ADRs (**H2.3**). The requirements are,

   (a) A low false positive rate.

   (b) To be efficient.

   (c) To be robust.

4. Evaluate the novel algorithm on the THIN database.

   (a) Compare the general signalling ability of the novel ADR signalling algorithm and the existing methods applied to the THIN database (**H2**).

   (b) Evaluate the novel ADR signalling algorithm's ability on the OMOP specified drug and health outcomes of interest (**H3**).

   (c) Generate new ADR signals (**H4**).

Chapters 3-6 focus on Objectives 1-4 respectively.

1.3 Thesis Organisation

The continuation of this thesis is organised as follows. Chapter 2 presents the literature review that is split into a pharmacovigilance section and a pattern recognition section. The pharmacovigilance section presents an overview of the current techniques and the recent advances. The Bradford-Hill causality considerations are discussed, as ADRs represent a causal relationship and the criteria may present the opportunity to distinguish ADRs from non-ADRs. The existing methods developed for different healthcare databases are summarised, with their connection to the Bradford-Hill causality considerations evaluated. The final part of the pharmacovigilance section focusses on the new initiatives currently taking place that aim to improve the way ADRs are signalled. The pattern recognition part presents the statistical learning theory view of supervised and semi-supervised learning. The main supervised and semi-supervised algorithms, that are used during the later chapters of the thesis, are summarised.

In Chapter 3, the benchmark for the ADR signalling ability of the THIN database is determined by applying a selection of the existing methods to the THIN database. As there is no perfect gold standard for signalling ADRs, two different comparisons were applied. The first comparison involved analysing all the possible drug-medical event pairs for a set of specific drugs and considering only the drug-medical event pairs listened as known ADRs on the website NetDoctor [176] to be ADRs and all other pairs to be non-ADRs. This enabled the evaluation of the methods when there is a large number of non-ADRs, but the comparison was limited due to the possibility of ADRs listed on NetDoctor being incorrect and due to unknown ADRs. The second comparison, termed the

specific comparison, only evaluated the drug-medical event pairs corresponding to ADRs listed on drug packaging or definitively known non-ADRs. The specific comparison enabled a more realistic evaluation and highlighted methods that are non-consistent.

In Chapter 4, attributes that are suitable inputs for a learning algorithm to distinguish ADRs and non-ADRs were proposed. The attributes included values from existing method, novel attributes derived from the Bradford-Hill causality criteria or novel attributes derived by considering the structure of the THIN database. The technique for extracting and cleansing the data are described and mathematical formula for calculating each attribute are presented.

In Chapter 5, the learning algorithm is developed and tentative results are presented. The first part of the chapter proposes a novel supervised technique that learns from a mixture of drugs and, once learned, can be applied to any drug. In the second part, a novel semi-supervised technique is presented that uses the limited number of known ADRs for a drug of interest to generate a model that is specific to the drug. Both the supervised and semi-supervised models are evaluated on a selection of drugs. The evaluation suggests that a supervised model trained on multiple drugs will outperform a semi-supervised model trained on a single drug. This is advantageous, as the supervised model can be trained on drugs that have been marketed for years and can be applied to newly marketed drugs that have limited toxicity knowledge.

In Chapter 6 the novel supervised algorithm is applied to more drugs and compared with a selection of existing methods using the specific comparison technique. The results showed that the novel supervised algorithm was often significantly better and had a better mean average precision (MAP) score and

lower false positive rate than existing methods. An additional evaluation was conducted by investigating the novel supervised algorithm's ability when considering the health outcomes of interest (HOIs) and drugs of interest (DOIs) specified by the OMOP. The evaluation showed that the novel supervised algorithm obtains a lower false positive rate than existing methods (0.16) and is able to signal a high proportion of definitively known ADRs. Therefore, the novel supervised algorithm has the potential to extract new pharmacovigilance knowledge and may help signal ADRs shortly after new drugs are marketed.

The final chapter of the thesis contains the conclusion that highlights the key results of the research and answers the research questions proposed in the introduction. Areas of future work are proposed, such as the modification of the algorithm to return quantitative information about the risk of each ADR signalled. The journal and conference contributions derived from this research are presented at the end of Chapter 7.

1.4 Contribution to Knowledge

This research has presented the first supervised and semi-supervised methods for signalling ADRs using LODs. New techniques for generating labels that are essential for supervised/semi-supervised algorithms have been presented and novel attributes that can distinguish between association and causation were proposed.

The research also highlighted the current limitations with evaluating the methods, as restricting the evaluation to a small number of definitively known drug-medical event pairs may prevent an accurate evaluation due to ignoring the numerous drug-medical event pairs that are associated due to confounding but

including more pairs into the evaluation may introduce error due to unknown ADRs.

This research has contributed to four journal papers (two published, one in print and one under review) and four conference papers. A full list of the journal and conference papers produced during this research is presented at the end of Chapter 7.

Chapter 2

Literature Review

> 'Prevention is the next frontier for pharmacovigilance, beyond simply generating alerts.'
>
> N. Moore [121]

2.1 Current Pharmacovigilance

2.1.1 Introduction

In 1961 a link was discovered between pregnant mothers ingesting the drug Thalidomide and then giving birth to infants with congenital malformations [167]. This widespread incident highlighted the importance of drug safety and prompted the start of systematic approaches to monitor the safety of marketed medications [29]. The research into medication safety is commonly referred to as pharmacovigilance. This involves the detection, assessment and prevention of ADRs for any marketed drug. The aim of pharmacovigilance is to identify ADRs, study relevant data and then investigate each ADR to assess risk factors. This knowl-

edge can then be used to help prevent ADRs that would otherwise lead to patient morbidity or mortality; helping to improve healthcare.

The process of identifying new ADRs involves signalling sets consisting of one or many drugs and an adverse event that may correspond to an ADR. There are different definitions for the term ADR signal in the context of pharmacovigilance but generally it means there is information to suggest a previously unknown causal relationship between some medication and an adverse event. The World Health Organization's (WHO) definition of an ADR signal is 'reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously. Usually, more than a single report is required to generate a signal, depending on the seriousness of the event and quality of the information' [53; 111]. Almenoff *et al.*(2005) interpret this as being able to 'view a signal as any information, qualitative or quantitative, that prompts further investigation on the relationship between a drug and an event' [1].

Once an ADR signal is generated, the medication and adverse event are studied further with more stringent statistical tests to confirm whether the signal is true, meaning there is sufficient evidence to confirm a causal relationship between the medication and the adverse event. Conversely, if there is not sufficient evidence, then the signal is false. In effect, ADR signalling is a way of filtering all the possible combinations of drug and adverse event pairs so that only the combinations that are most likely to correspond to ADRs remain to be investigated further. This is important as it is not possible to efficiently investigate the thousands or even millions of possible combinations of drugs and suspected ADRs in fine detail.

Overall, the process of identifying an ADR (a causal relationship between a drug and medical event) requires three steps [136];

1. **Signal generation/detection**- this step involves analysing all drug-medical event pairs representing a possible ADR and highlighting the ones that are most suspicious.

2. **Signal refinement**- after signals are generated in step 1 for some drug-medical events pairs these drug-medical event pairs are actively surveyed to look for more evidence that they may correspond to an ADR.

3. **Signal evaluation**- this is when a single in depth investigation (formal epidemiological study) is performed to determine if there is causality between a drug and a medical event that has been signalled in step 1 and refined in step 2.

Early pharmacovigilance depended on professionals manually investigating hard copies of reports detailing suspected ADRs. These professionals would then identify commonly occurring suspected ADRs or highly noxious suspected ADRs as signals [113; 145]. The limitations with this methodology was that collaboration was difficult prior to the World Wide Web so the reports were only collected from a segment of the population and less obvious ADRs may never have been suspected and reported or may have been difficult to identify. With the advances in technology enabling rapid communication between borders and helping pool large quantities of data together, many of the original limitations are beginning to disappear [198]. Using the large collections of electrically stored data, we are now presented with the opportunity to apply data mining methods and generate ADR signals more efficiently [35], as less time is required before there is a

sufficient number of ADR incidences reported to enable the signal generation [158].

The majority of existing pharmacovigilance methods that use large databases for ADR signal detection are applied to the SRS databases. The SRS databases are readily available electronic databases that contain a collection of voluntary reports of suspected ADRs, often containing millions of reports. This type of database has been used to successfully signal many ADR signals, but the signals cannot be considered definitive [119]. There are also well documented limitations with using SRS databases for ADR signalling, due to these databases relying on people recognising and reporting suspected ADRs [63; 79; 168]. It has been suggested that these limitations may prevent the detection of rare ADRs [79] or lead to delays in generating ADR signals [93]. The standard procedure for automating the generation of ADR signals in SRS databases relies on calculating a measure of disproportionality corresponding to how often the adverse event is reported after a specific drug compared to a baseline determined by how often it is reported after any drug within an SRS database [119]. As there is no gold standard for ADR signal detection, each country tends to have a different preference for the choice of disproportionality method applied to his SRS database. The main disproportionality methods applied to the SRS databases and their limitations are detailed in Chapter 2.1.3.

Recently the LODs have caught the attention of pharmacovigilance researches as it offers a unique perspective for ADR signal generation and is starting to become more readily available [180]. The LODs suitable for pharmacovigilance contain timestamped medical records and timestamped prescription records for patients over large periods of time. Rather than relying on people suspecting

ADRs like the SRS databases, potential ADRs can be inferred using temporal relationships between the medical and prescription records for a patient. This may enable the detection of rare ADRs or ADRs with a high background rate [180] that can not be identified by mining the SRS databases. Additionally, as generating ADR signals by mining the LODs does not require people noticing potential ADRs, it may be possible to generate ADR signals earlier than by mining the SRS databases. In Chapter 2.1.4, the current ADR signalling methods developed for the LODs are described, along with their limitations.

The continuation of the pharmacovigilance section of the literature review includes a section summarising causality, and the frequently applied Bradford-Hill causality considerations [19]. This is followed by a description of the current methodologies for detecting ADRs by mining the SRS databases and the new advances into mining ADRs using LODs. The final section summarises the current pharmacovigilance initiatives and highlights how this field of research may change with the integration of multiple electronic healthcare databases.

2.1.2 Causality

The definition of a statistical association is 'a relationship between two measured quantities that renders them statistically dependent' [183] . Whereas the term causality is defined as 'a relationship between two events, the cause (or incidence) event and effect (or consequent) event, where the effect event is dependent on the cause event' [169]. Therefore a causal relationship is also an association but not all associations are causal. It is clear that an ADR represents a causal relationship, as the adverse event is a result of a patient ingesting a specific drug and would

not have occurred if the patient did not have the drug. The first step in current ADR discovery, namely signal generation, finds drug-medical event pairs that are associated and the later two steps, signal refinement and evaluation, aim to determine if the found association is also a causal relationship.

A common method to assess causality between an antecedent and consequence is to use the Bradford-Hill causality considerations [19] that proposes nine factors that need to be considered,

1. **Strength**- how much do the antecedent and consequence appear to be associated? A high association would suggest causation however a low association does not mean there is no causation.

2. **Consistency**- has the relationship been observed in different patients and situations? (In the context of ADRs, has it been reported in multiple patients and databases?).

3. **Specificity**- is the relationship specific (e.g., there are few other associations containing the antecedent or consequence). This factor has limitations as many ADRs are the result of multiple causes. An alternative interpretation is whether the population experiencing the relationship is specific (e.g., old, young or female).

4. **Temporality**- the order of the antecedent and consequence (e.g., did the medical event make the patients more prone to the drug or did the drug cause the medical event?).

5. **Biological Gradient**- is there an increasing monotonic relationship between the frequency/amount of the antecedent and the frequency of the

consequence (e.g., does a higher dosage of the drug increase the medical event frequency?).

6. **Plausibility**- does it make sense? However, this is not a necessary feature as plausibility depends on current knowledge and even the improbable could be true.

7. **Coherence**- does the relationship conflict with known facts? (e.g., Do we know drug-medical events pairs that are definitely not ADRs?).

8. **Experimentation**- does changing the antecedent change the consequence? (e.g., Does the medical event start when the drug starts and stop when it stops?).

9. **Analogy**-are there similarities with known causal relationships (e.g., does the ADR exist for a similar drug)?.

The more Bradford-Hill causality factors covered by a method, the more likely it is to correctly identify causal relations and therefore identify ADRs. In Chapters 2.1.3.2 and 2.1.4.3 the range of Bradford-Hill causality factors considered by each of the existing ADR signalling methods are determined.

2.1.3 Spontaneous Reporting Databases

2.1.3.1 Overview

The SRS databases were one of the first resources to contain vast quantities of pharmacovigilance data and enable an aggregated analysis [152]. Their presence in the field of pharmacovigilance has aided the discovery of many ADRs [106],

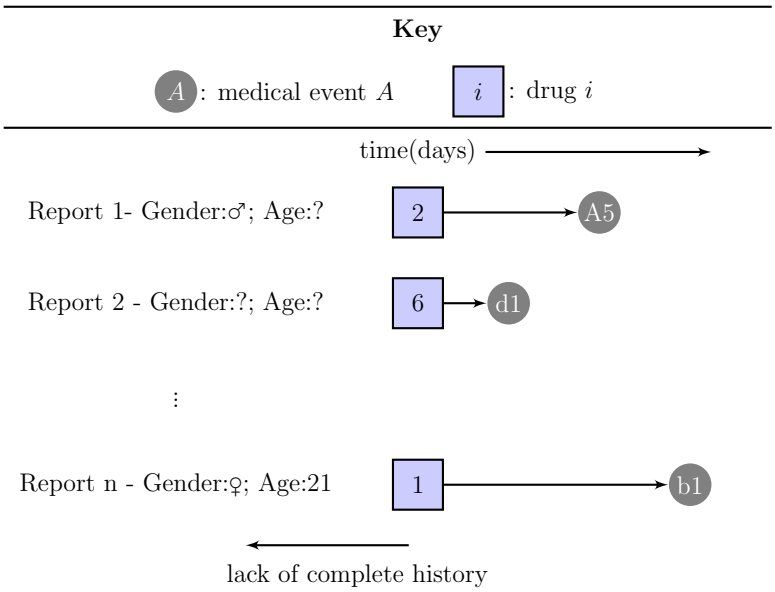


Figure 2.1: Illustration of data contained in SRS databases.

but their application is limited [63; 172]. The databases contain linked drug and medical event records. Each link represents that the drug was a suspected cause of the medical event. In additional to the linked drug and medical records, there are also details specifying information about the patient that experienced the suspected ADR. An illustration of the data contained in SRS database is presented in Figure 2.1, where drugs are represented by squares and medical events are represented by circles. An example of the database design for an SRS database can be seen in Figure 2.2. The records in the database are submitted voluntarily by medical practitioners or the general public [118]. Two common examples of SRS databases are the Food and Drug Administration (FDA) Adverse Event Reporting System (AERS) [77; 184] in the USA and the Yellow Card Scheme SRS [118] run by the Medicines and Healthcare products Regulatory Agency (MHRA) and the Commission on Human Medicines (CHM) in the UK.

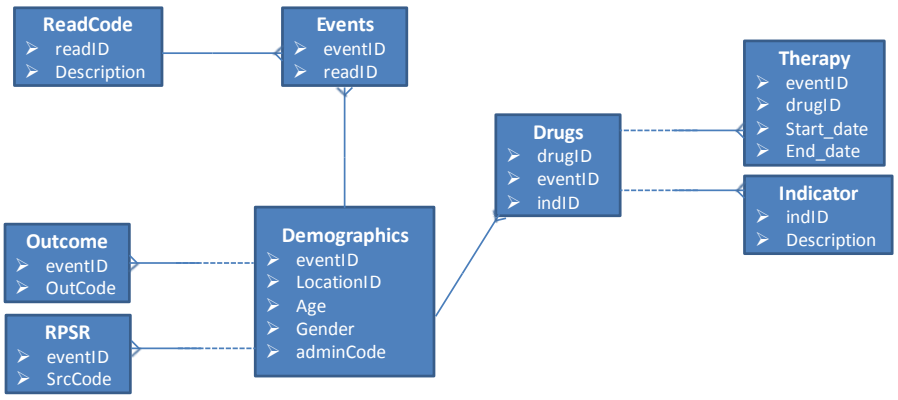


Figure 2.2: An example entity relationship diagram for an SRS database based on the FAERS database.

The general process involved in reporting a suspected ADR into an SRS database is for a patient or doctor to fill out a form detailing the drug/drugs prescribed, the adverse event experienced, some information about the patient and information about the person making the report. An example of a typical SRS report submitted online via the yellow card scheme in the UK can be seen in Figure 2.3. The majority of fields in the form are not required and the entries are not validated during submission. This causes limitations as it is common for SRS databases to contain missing or incorrect data [137]. It is also known that SRS databases suffer for bias reporting [173], especially under reporting [79]. Another issue is underascertainment, when the ADRs is not noticed (e.g., ADRs corresponding to medical events with a high background rate or rare ADRs may never be suspected) [173].

The SRS methods determine the association strength between a drug-medical event pair and pairs with a high association are signalled. The generated signals require further analysis as association does not imply causation [1]. This has

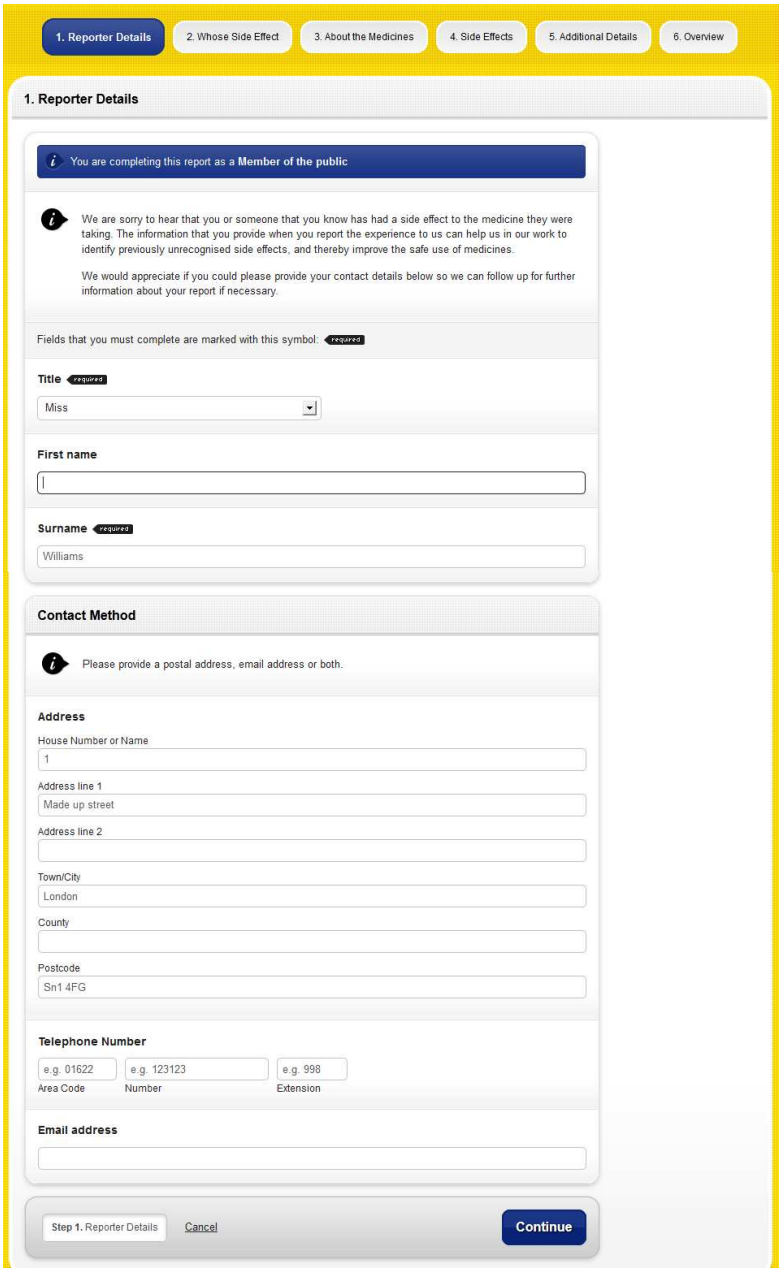


Figure 2.3: The online form for submitting suspected ADRs via the Yellow Card Scheme in the UK.

prompted researches to state that the SRS methods are 'initial filters' for identifying ADR associations [8] and are not capable of generating definitive signals. Once the SRS methods generate a signal, it is then refined and finally evaluated. This means the causal relationship is not confirmed until much later in time than when the original signal occurred.

The SRS databases generally have a fixed point in time perspective, as limited past and present medical knowledge for each patient is known [12], the lack of historical data is illustrated in Figure 2.1. The actual rate that a drug is prescribed and the rate that a medical event occurs is unknown [172], as SRS databases only contain data on the drug prescriptions that may have resulted in an ADR. Consequently, the SRS methods estimate the baseline rate that a medical event occurs by finding out how often the medical event is reported with any drug in the database. Medical events that are reported disproportionally more often with the drug of interest compared to all the other drugs in the database are then ranked highly as suspected ADRs. The methods make use of a contingency table, see Table 2.1, summarising the number of reports that contain (or do not contain) the drug and event of interest. Each method estimates the baseline rate differently, by using different combinations of the values in Table 2.1. Unfortunately, the estimation of the background rate, by using other drug reports, can limit the signals that are generated [76] and prevent some ADRs (e.g., those with a high background rate) being identified. In addition, both over-reporting and under-reporting can lead to skewed estimates for the background rates and influence the signalling ability.

Initially, the disproportionality methods relied on calculating measures linked to standard epidemiology statistical values such as the Reporting Odds Ratio

Table 2.1: A sample contingency table used by the disproportionality methods applied to the SRS databases.

	Event Y	Other Event	Total
Drug X	a	b	a+b
Other Drug	c	d	c+d
Total	a+c	b+d	a+b+c+d

Table 2.2: The different SRS methods and the measures they implement to calculate the association between a drug-medical event pair. [1]

Method	Measure	Probabilistic Interpretation [75]	Approach
ROR	$\frac{a/b}{c/d}$	$\frac{\text{P(AE\|Drug)/P(notAE\|Drug)}}{\text{P(AE\|notDrug)/P(notAE\|notDrug)}}$	F
PRR	$\frac{a/(a+b)}{c/(c+d)}$	$\frac{\text{P(AE\|Drug)}}{\text{P(AE\|notDrug)}}$	F
NPRR	$\frac{a/(a+c)}{b/(b+d)}$	$\frac{\text{P(Drug\|AE)}}{\text{P(Drug\|notAE)}}$	F
BPCNN (IC)	$log_2(\frac{a(a+b+c+d)}{(a+c)(a+b)})$	$log_2(\frac{\text{P(AE\|Drug)}}{\text{P(AE)}})$	B
EBGM (RR)	$\frac{a(a+b+c+d)}{(a+c)(a+b)}$	$\frac{\text{P(AE\|Drug)}}{\text{P(AE)}}$	B

(ROR) [9] and Proportional Reporting Ratio (PRR) [185]. In [196] the authors propose a novel PRR (NPRR) method, that takes a slightly different perspective, and they suggested that both the PRR and NPRR should be used to generate a signal. More recently, methods have been implemented that are based on artificial neural networks, such as the Bayesian Propagation Confidence Neural Network (BPCNN) [10], or Bayesian modelling, such as the Empirical Bayesian Geometric Mean (EBGM) [50]. Table 2.2 summarises the different methods and displays their probabilistic derivations. The SRS signal generation methods are split between frequentist statistical approaches (ROR, PRR) and Bayesian statistical approaches (EBGM, BPCNN). The frequentist statistical methods assume that the parameters for a model are fixed and they consider that the data comes from

[1] In the approach column, F represents frequentist and B represents Bayesian.

a repeatable random sample. These methods do not require prior knowledge of a model and are computationally cheap. It follows that the advantage of frequentist methods for signal detection is that they are fast, which is an important factor due to the large quantities of data available. Conversely, the Bayesian statistical methods assume that the data are fixed and the parameters are unknown but described by a probabilistic distribution. These methods require some prior knowledge and can be computationally costly. The advantage of using Bayesian methods for signal detection is that, due to the parameters being non-fixed, they can adapt over time when changes in the drug prescription habits may differ, such as when doctors change the prescription rates of drugs or prescribe drugs to patients for a non-standard indication.

The methods all have signalling criteria, see Table 2.3. The frequentist methods generate a signal for a drug-medical event pair when there are three or more case reports and the lower 95% confidence interval is greater than one. Their standard errors, displayed in Table 2.4, are estimated using the woolf logit method [205], a method that approximates the distribution of the ln(ROR) and ln(PRR) as being normal. The EBGM generates a signal for a drug-medical event pair when the lower bound of the 90% credibility interval, EB05, is greater than two. The BPCNN signals a drug-medical event pair when its IC value minus two standard deviations changes from negative to positive.

2.1.3.2 Causality

The SRS methods all work out the association strength between a drug and medical event based on the disproportionality measure. As the SRS databases sometimes contain the patient details such as age and gender it is possible for

Table 2.3: The signalling criteria for the different SRS methods [50; 185] .

Method	Signal Criteria	Shrinkage
ROR	$exp[ln(ROR) - 1.96SE(ln(ROR))] > 1$	No
PRR	$exp[ln(PRR) - 1.96SE(ln(PRR))] > 1$	No
NPRR	$exp[ln(NPRR) - 1.96SE(ln(NPRR))] > 1$	No
EBGM	$EB05 \geq 2$	Yes
BPCNN	$IC - 2SD > 0$	Yes

Table 2.4: The standard errors for the frequentist methods [185; 196].

Method	Standard Errors
ROR	$SE(ln(ROR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
PRR	$SE(ln(PRR)) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$
NPRR	$SE(ln(NPRR)) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$

the SRS algorithms to deal with the specificity criteria, but none of the existing methods does and the problem of missing values may make this difficult. The methods do not cover the consistency criteria when they are only applied to one SRS database and they do not deal with the biological gradient as they do not take into consideration the dosage of the drug. The SRS methods estimate the background risk of a medical event based on all other drugs rather than restricting themselves to estimating the risk of the medical event based on similar drugs, therefore they do not consider the analogy criterion. Due to their restricted perspective, they cannot cover the experimentation factor as this requires observing what happens when the drug stops and starts. The temporality, plausibility and coherence criteria are indirectly covered as people should only submit a report when a drug is suspected to have caused an ADR, and any suspected ADR would have occurred after the drug is taken and must be plausible and coherent otherwise it would not be suspected. However, people may make mistakes when

Table 2.5: The Bradford-Hill causality considerations covered by each method. [1]

Criteria	ROR	PRR	EBGM	BPCNN
Strength	✔	✔	✔	✔
Consistency	×	×	×	×
Specificity	×	×	×	×
Temporality	•	•	•	•
Biological Gradient	×	×	×	×
Plausibility	•	•	•	•
Coherence	•	•	•	•
Experimentation	×	×	×	×
Analogy	×	×	×	×

reporting a suspected ADRs or may not know information that would otherwise make them reconsider that the medical event is a suspected ADR.

The criteria that could be covered by SRS methods, but are not currently, are the consistency, analogy and possibly the specificity. The consistency criteria could be covered by using other SRS databases as a cross reference to see if there is evidence in other databases for the signals generated by the SRS methods. The analogy could be covered by comparing a drug of interest against drugs in the same family and using knowledge about existing ADRs for similar drugs and the specificity could be covered by comparing the drug and event disproportionalities between different groups of the population, such as the old or young. This information is summarised in Table 2.5.

2.1.3.3 Limitations

The main limitations with signal generation using the disproportionality methods applied to SRS databases are due to database issues. The databases are known to contain missing or duplicated data and suffer from inconsistent reporting such as

[1] • represents indirectly covered.

under-reporting or over-reporting for new drugs or more serious adverse events [9]. It is often common for the SRS databases to be plagued by inconstancies due to changes in medical terminology over time or variance in the level of detail recorded for the medical event depending on the person making the report [9]. Bate *et al.* (2009) state that 'disproportionality methods do not estimate reporting rates', as the reporting rate calculation requires the knowledge of drug usage and this is not contained in the SRS databases [9]. The effect of this is that specific adverse events will not be found, such as when a drug causes all events to increase or when an adverse event is common for many drugs. The two major consequences of the under-reporting are that there may be a large time lag between when a rare ADR is first reported and when it is signalled, or it is possible that rare ADRs may never be detected. One retrospective study found that 19.6% of known ADRs were signalled by the PRR after other pharmacovigilance methods and 26.9% of known ADRs had not been signalled during the study period [3]. The SRS methods cannot be used for signal refinement or evaluation due to the limitation of not knowing the actual background rates that medical events occur or drug are prescribed. Therefore, it is not possible to develop a method capable of definitively identifying ADRs that only uses the SRS databases, instead, other types of databases are required for the signal refinement and evaluation once signals have been generated by mining the SRS databases.

2.1.3.4 Summary

Mining the SRS databases has aided new ADR discoveries, but the signals generated by mining the databases require further evaluation and the majority of signals do not lead to ADR discovery [172]. A recent study provided evidence

to suggest that limitations due to how the SRS data are collected may make it difficult for the disproportionality methods to identify ADRs with a high background rate [70]. Furthermore, as a consequence of the the limited perspective of the SRS databases, they cannot be used to quantify ADR risks [119], nor can they be used to identify risk factors.

2.1.4 Longitudinal Observational Databases

2.1.4.1 Introduction

The LODs are databases containing temporal medical data [12] on thousands or millions of patients, often spanning over many patient years. An example of an LOD is The Health Improvement Network (THIN) database, see appendix A, that is an electronic database containing the data stored in over 500 UK general practices [65]. The data consists of patient details such as their year of birth, gender, family links and timestamped medical and prescription records. It has been found to be a suitable representation of the UK [15] and it is not common to find duplicated or missing data due to validation procedures. Researchers have assessed the validity of using the THIN database for pharmacovigilance by investigating whether known associations can be found using the data and concluded that its use is valid [105]. The database contains records of every medical event for a patient that the doctor has been informed of, as the data is extracted directly from the local GP databases and doctors must record all the relevant medical details each time a patient visits [85]. Unfortunately underreporting is still possible in these types of databases, as some drugs can be bought rather than being prescribed and patients may not inform their doctor of all the

medical events that they experience. It is also possible for LODs such as the THIN database to have inconsistencies in data recording between practices [78]. The THIN database also has issues with patients changing practices, as each patient is given an anonymous ID within their practice but when they change practices they will receive a new ID, and there are no links between the two IDs to identify them as representing the same patient [104].

The LODs offer a unique perspective for discovering ADRs as, unlike the SRS databases, they do not have direct links between drugs and medical events that are potential ADRs [161] but potential ADRs can be inferred using the temporal information. For example, if investigating ADRs that occur immediately after taking a drug, all the medical events that occur within 30 days of taking the drug can be flagged as potential ADRs. The advantage of generating ADR signals using the LODs compared to the SRS databases are they contain patients' medical histories and include patients that did not experience adverse events after taking a drug [71]. Therefore they contain the background rates that a drug is prescribed or a medical event occurs [216] and are less prone to bias reporting due to not relying on voluntary reports. As the LODs are not restricted to finding ADRs that occur shortly after taking a drug, they could be used to find ADRs that are not present till many years after taking a drug. Furthermore, the vast quantities of data contained in LODs makes them more suitable for detecting drug-drug interaction ADRs or child specific ADRs. Figure 2.4 illustrates the data contained in LODs, and shows that the drug-medical event pairs that are potential ADRs can be found by investigating the $[t_0, t_1]$ period around each prescription. Another advantage of the LODs is that they have frequently been used for signal refinement and evaluation [31], so all three steps of detecting an ADR can be implemented

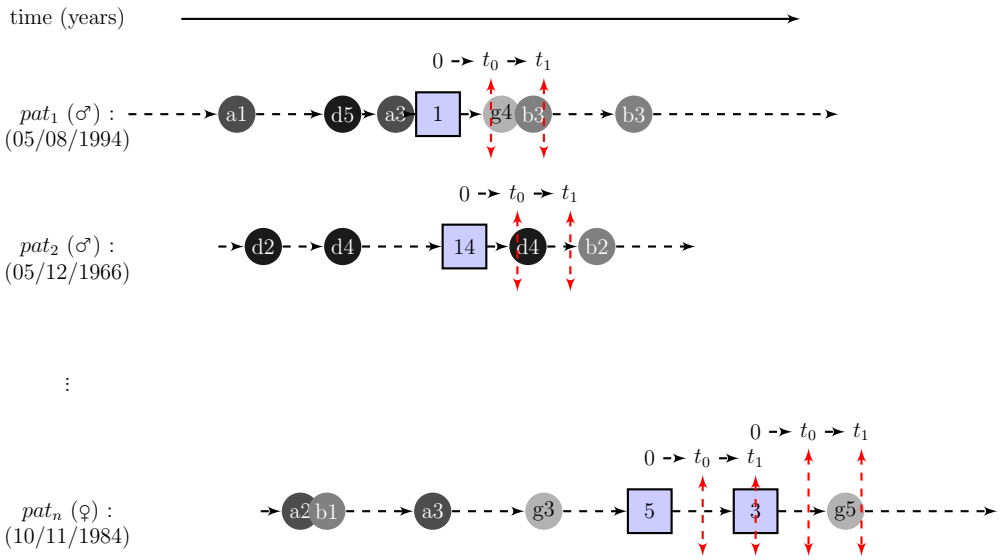


Figure 2.4: Illustration of patients' longitudinal data contained in the THIN database.

on a single LOD, making it possible to develop an efficient algorithm capable of definitively detecting ADRs. Although, there may be issues with performing signal evaluation on the same data used to generate the signal.

Numerous approaches have been suggested to signal ADR using LODs [26; 83; 90; 128; 216], but there is currently no algorithm that has been developed specifically for the THIN database . The methods tend to calculate a measure of association between each medical event and drug. This is calculated by comparing the risk of the medical event for the drug taking population within a defined time interval after the drug is prescribed with the risk of the medical event in some substituted population. These methods are based on the counterfactual theory of causality, where the observed risk of the medical event in the drug taking population is compared with the risk that would have been observed had the patients not taken the drug [122]. Once the patients take the drug, the second

situation (i.e., patients not taking the drug) is counterfactual and unobservable, so an observable substitution is use instead to approximate the second risk. If the substitution does not match the counterfactual, then confounding is introduced and the measure of association differs from the measure of causation [66].

An example of the counterfactual theory of causation is presented in Figure 2.5. It can be seen that the patient 1 given treatment 1 experienced medical event A but would not have experienced it if treatment 0 was given, so medical event A was caused by being given treatment 1 rather than treatment 0. However, it is impossible to observe patient 1 taking only treatment 1 and only treatment 0 at the same time, therefore an observable substitution is used to estimate causality. Association is determined by observing patient 2 taking treatment 0 and comparing the outcome with patient 1 taking treatment 1. Unfortunately, as the patients are different, the observed outcome over $[t_0, t_1]$ for patient 2 taking treatment 0 is different to what would have been observed for patient 1 given treatment 0 and the substitution comparison indicates that both medical event A and medical event B are associated with treatment 1. However, only medical event A is caused by treatment 1, the medical event B association is due to confounding introduced by the substitution.

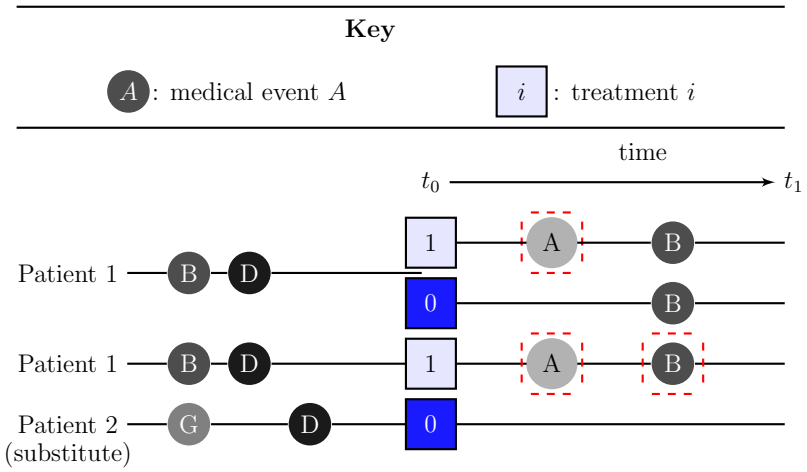


Figure 2.5: Illustration of the counterfactual theory of causaltion.

2.1.4.2 Methods

Disproportionally Methods

The disproportionality methods, such as modified SRS [216] and Temporal pattern Discovery (TPD) [128], compare the risk during the time interval $[t_0, t_1]$ centred around the drug of interest prescriptions with the risk during the time interval $[t_0, t_1]$ centred around all drug prescriptions, so the substituted population is the patients taking any drug. This is illustrated in Figure 2.6. The TPD also looks for temporal changes in the measure of association, as this reduces the effect of confounding by indication (i.e., when differences arise between the patients taking the drug and those not taking the drug), as illustrated in Figure 2.7. Justification for using all other drug reports as a substitution, but keeping the same time interval of interest, is that medical events are not reported uniformly over time [128], and it is common for the majority of medical events to be reported shortly after a prescription. By investigating the same period of time relative to the prescription, the potential bias caused by non-uniform reporting

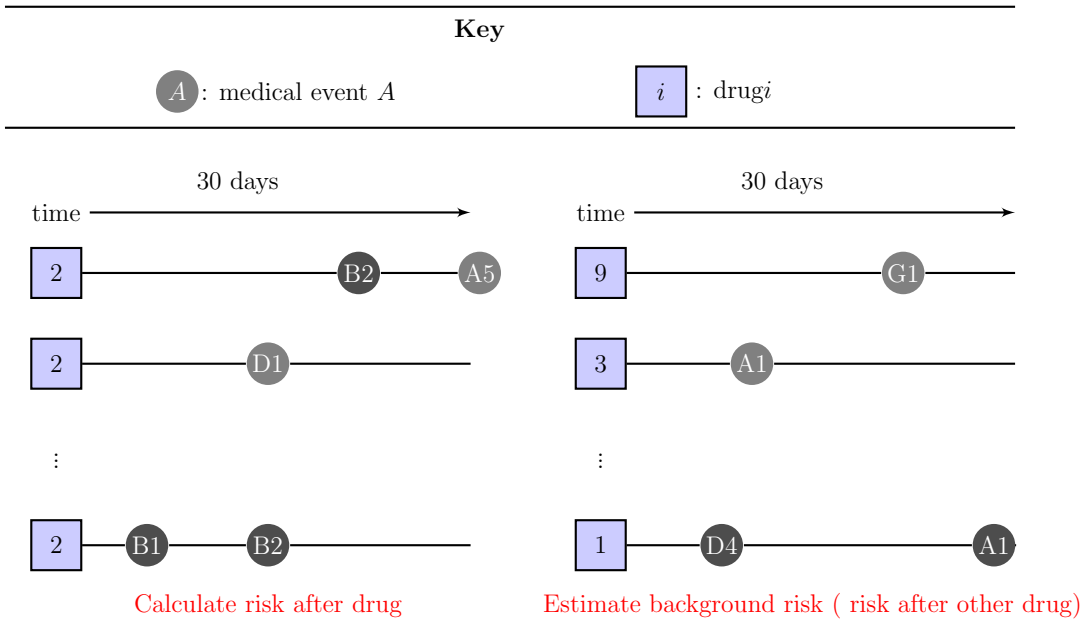


Figure 2.6: Illustration of the disproportionality methods.

is removed.

To apply the standard SRS methods, described in Chapter 2.1.3, the contingency tables need to be determined using the LOD data. In [216], the authors presented three different proposals for calculating the contingency tables for a specific drug x and medical event y using LODs. The spontaneous reporting system (SRS) and modified-spontaneous reporting system (modified-SRS) approaches performed similarly and both outperformed the distinct patient approach. Referring to the set time period after the drug of interest is prescribed as the drug hazard period, the SRS approach calculates the a-d values in Table 2.1 as, a is the number of distinct times event y occurs during any x hazard period, b is the number of distinct times any non–y event occurs during any x hazard period, c is the number of distinct times event y occurs in any non–x hazard period and d is the number of distinct times any non–y event occurs within any non–x hazard

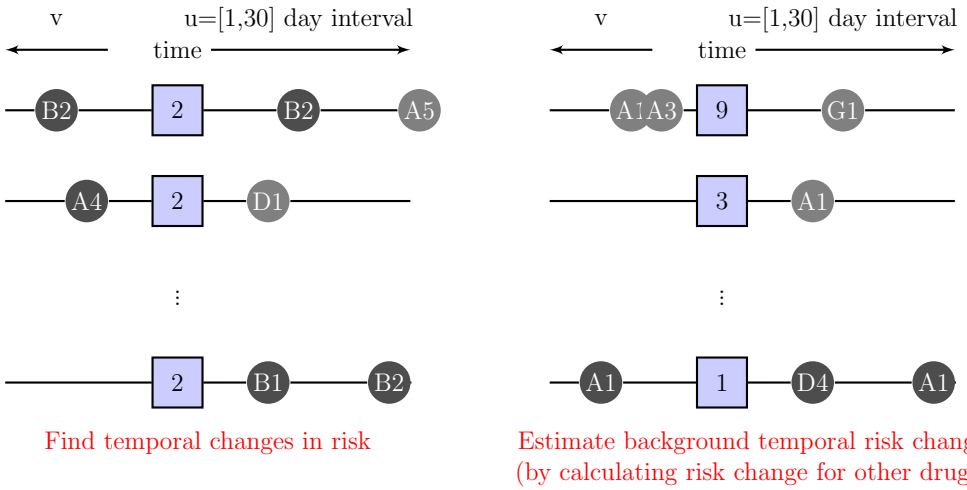


Figure 2.7: Illustration of the TPD method.

period. The modified-SRS approach is similar but considers the prescriptions that do not have medical events recorded. Therefore, b becomes the number of distinct times any non–y event occurs during any x hazard period plus the number of x hazard periods that have no medical event recorded and d becomes the number of distinct times any non–y event occurs within a non–x hazard period plus the number of non–x hazard periods that have no medical event recorded plus the number of distinct times non–y events are reported outside of a hazard period.

The TPD method [128] compares the amount of patients that have the first prescription of drug x in thirteen months followed by event y within a set time t relative to the expected number of patients if drug x and event y were independent. The background rates that a medical event occurs is calculated based on how often it occurs within the hazard period for any drug. Letting,

$n_{.y}^{t}$ denote the number of patients that are prescribed any drug for the first time in 13 months and have event y within time t.

$n^t_{x.}$ denote the number of patients that have drug x for the first time in 13 months and are registered for any period over time t

$n^t_{..}$ denote the number of patients that have any other drug for the first time in 13 months and are registered for some period over time t.

n^t_{xy} denotes the number of patients that have drug x for the first time in 13 months and event y occurs within time t after.

The expected number of patients that have drug x and then event y in a time period t is then,

$$E^t_{xy} = n^t_{x.} \frac{n^t_{.y}}{n^t_{..}} \tag{2.1}$$

If for a given drug, the event occurs more than expected, the ratio between the observed and expected will be greater than one. By taking the log_2 of the ratio, a positive values suggests an interesting association between a drug and event. Modifying the equation to prevent the problem of rare events or drugs resulting in a small expectation that can cause volatility, a statistical shrinkage method is applied.

$$IC = log_2 \frac{n^t_{xy} + 1/2}{E^t_{xy} + 1/2} \tag{2.2}$$

The shrinkage adds a bias for the IC towards zero when an event or drug is rare. The credibility intervals for the IC are the logarithm of the solution to equation 2.3 with $q = 0.025$ and $q = 0.975$.

$$\int_0^{\mu_q} \frac{(E^t_{xy} + 1/2)^{n^t_{xy}+1/2}}{\Gamma(n^t_{xy} + 1/2)} u^{(n^t_{xy}+1/2)-1} e^{-(n^t_{xy}+1/2)} du = q \tag{2.3}$$

The above can find possible drug and event associations of interest for a given

t, however, the authors suggest that general temporal patterns can be found by comparing the IC of two different time periods. The follow-up period of primary interest is denoted by u and the control time period by v. This removes event and drug relationships that just happen to occur more in certain sub-populations. The different between the IC for both time periods is,

$$log_2 \frac{n^u_{xy}}{E^u_{xy}} - log_2 \frac{n^v_{xy}}{E^v_{xy}} \tag{2.4}$$

re-arranging and adding a shrinkage term gives,

$$IC_\Delta = log_2 \frac{n^u_{xy} + 1/2}{E^{u*}_{xy} + 1/2} \tag{2.5}$$

where

$$E^{u*}_{xy} = \frac{n^v_{xy}}{E^v_{xy}} . E^u_{xy} \tag{2.6}$$

As it was observed that medical events related to the cause of the drug are often assigned a high IC value after the prescription but also prior to the time the drug is prescribed, the TPD algorithm includes a filter thats ignores medical events that have a higher IC value on the day of prescription or a month before the prescription relative to the month after the prescription.

Methods that calculate association tend to suffer from confounding as association does not imply causation, so many of the medical events signalled due to a high association value may not be ADRs. One method that has been presented to counteract the problem of confounding is the ROR Regression (RORR) method [72]. The RORR effectively filters the drugs that are signalled as ADRs by the

ROR by determining whether the association may be due to confounding. The method applies two regression models, the first model does not consider the effect of covariates, letting y represent the medical event, and x_1 represent the drug, then the log odds of medical event y is,

$$log(\frac{P(y|x_1)}{1-P(D|x_1)}) = b_0 + b_1 x_1 \tag{2.7}$$

where b_0 is the background log odds ratio of medical event y. The second model considers the effects of the covariates, $x_i, i > 1$, and the log odds of y is calculated as,

$$log(\frac{P(y|x_1, x_2, ..., x_k)}{1-P(D|x_1, x_2, ..., x_k)}) = b_0 + \sum_{i=1}^{k} b_i x_i \tag{2.8}$$

For each drug with a high ROR, the regression model only considering the drug, equation (2.7), and the regression model considering all the covariates, equation (2.8), are both applied and drugs that have similar b_1 values for both models are considered to be causes of medical event y. Unfortunately, its application on the THIN database is currently limited due to the requirement of choosing the appropriate covariates for each signal. This requires manual expert input for each signal, which would be time consuming.

Sequential Pattern Methods

Methods based on sequential pattern mining include Mining Unexpected Temporal Association Rules given the Antecedent (MUTARA) [91] and Highlighting UTARs Negating TARs (HUNT) [90]. These methods calculate the standard sequential patterning mining measure known as leverage [134] that subtracts the expected proportion of all sequences that contain the drug followed by the medical event within a defined time interval from the observed proportion. The expecta-

tion is derived by calculating the risk within a randomly selected time interval for the population of patients never prescribed the drug. In effect, this is similar to a retrospective cohort study as the cohorts are the patients exposed or non-exposed to the drug.

The authors of MUTARA and HUNT refer to the patients prescribed the drug as users and patients never prescribed the drug as non-users. Both methods first restrict their attention to subsequences of the user and non-user sequences. For each user sequence, the T_h constrained subsequence of interest is the subsequence of length T_h days starting from the day the drug is first prescribed. The value of T_h differs between users depending on whether the user has a repeat prescription within T_e days after the first prescription. If the user does not have a repeat prescription within T_e days of the first prescription then $T_h = T_e$, whereas if the second prescription of the drug occurs s days after the first prescription where $s \leq T_e$ then $T_h = s + T_e$. For each non-user, the T_c constrained subsequence of interest is a subsequence of length T_c days that is randomly chosen from the non-user's sequence. An illustration of this can be seen in Figure 2.8.

Defining *tot* as the number of users and non-users, the $supp(x \xrightarrow{T} y)$ is defined as the number of user T_h constrained subsequences containing the medical event y divided by *tot*, the $supp(x \xrightarrow{T})$ is the number of users divided by *tot* and $supp(\xrightarrow{T} y)$ is the number of user T_h constrained subsequences that contain the medical event y divided by *tot* plus the number of non-user T_c constrained subsequences that contain the medical event y divided by *tot*. The leverage is calculated as,

$$Leverage = supp(x \xrightarrow{T} y) - supp(x \xrightarrow{T}) \times supp(\xrightarrow{T} y) \tag{2.9}$$

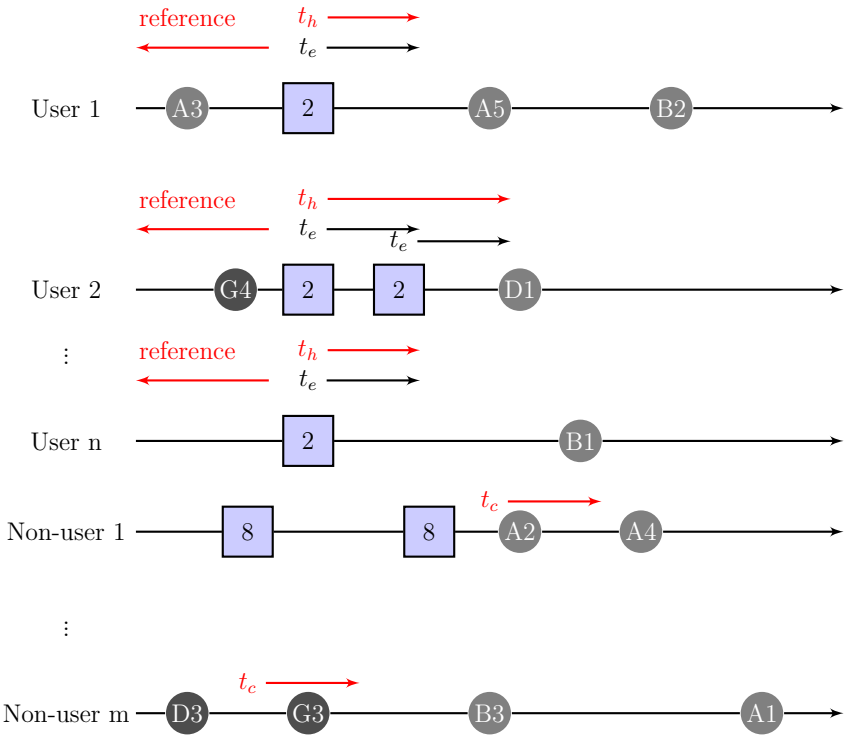


Figure 2.8: Illustration of the MUTARA and HUNT methods.

In addition to calculating the standard leverage, a new measure called unexpected-leverage is also calculated. The unexpected-leverage (unexlev) makes use of a user's history to filter repeated medical events from the users's T_h constrained subsequence as these are 'predictable' and unlikely to be ADRs. This is done by investigating a reference period prior to the first prescription within the user's sequence and filtering medical events from the user's T_h subsequence if they occurred during the reference period. Defining $\text{supp}(x \overset{T}{\hookrightarrow} y)$ as the number of users who's T_h constrained subsequence contains medical event y but who do not have medical event y within the reference period divided by tot and $supp(\overset{T}{\hookrightarrow} y)$ as the total of the number of users whose T_h constrained subsequence contains medical event y but who do not have medical event y within the reference period plus the number of non-user T_c constrained subsequences that contain the medical event y all divided by tot, the unexpected leverage is calculated as,

$$unexlev = supp(x \overset{T}{\hookrightarrow} y) - supp(x \overset{T}{\rightarrow}).supp(\overset{T}{\hookrightarrow} y) \tag{2.10}$$

MUTARA returns medical events ordered by $unexlev$ and HUNT returns medical events in descending order of the ratio between the leverage rank and the unexpected-leverage rank,

$$RankRatio = \frac{\text{medical event rank based on leverage}}{\text{medical event rank based on unexpected-leverage}} \tag{2.11}$$

Other Methods

Other methods for signalling ADRs using LODs that have been proposed include fuzzy logic methods [89], calculating the log likelihood over time [26], applying a

Table 2.6: The different LOD ADR signalling algorithms and the causality criteria each of them covers. [1]

Criteria	MUTARA	HUNT	TPD	Modified SRS	RORR
Strength	✔	✔	✔	✔	✔
Consistency	✗	✗	✗	✗	✗
Specificity	✗	✗	✗	✗	*
Temporality	✔	✔	✔	✗	✗
Biological Gradient	✗	✗	✗	✗	*
Plausibility	✗	✗	✗	✗	✔
Coherence	✗	✗	✗	✗	✗
Experimentation	✗	✗	✗	✗	✗
Analogy	✗	✗	✗	✗	✗

sequential version of the self controlled case series [83] or adapted epidemiology based approaches, see Chapter 2.1.5. These methods tend to suffer from confounding effects and are likely to have a high false positive rate. However, it is worth noting that the self controlled case series is resilient to any fixed in time confounding. Very few of these methods have been implemented on a range of LODs, so their robustness is unexplored.

2.1.4.3 Causality

The LOD ADR signalling algorithms all cover the strength criteria as they calculated the dependancy of the occurrence of a medical event on the occurrence of a drug being prescribed. The filtering in the MUTARA/HUNT and the TPD algorithms means they cover the temporality criteria as medical events that occur before the drug are generally filtered. The modified SRS and RORR algorithms do not apply a filter, so they do not cover the temporality criteria. In effect, the RORR covers plausibility by filtering out drug-medical event pairs that are asso-

[1] * means that the factor could be incorporated but is currently not.

ciated due to other causes, so the remaining drug-medical event pairs are more plausible ADRs. Additionally, it would be possible to include dosage and personal attributes into the regression model used by the RORR, so the specificity and biological gradient could be included. The other causality criteria are not covered by the LOD algorithms. This is likely to be the reason why the existing LOD algorithms frequently signal medical events linked to the cause of taking the drug or medical events that are just common in the drug taking population.

2.1.4.4 Limitations

The LOD databases have presented the opportunity to signal ADRs without the limitations associated with the SRS databases, but research has shown they have their own limitations [100; 131]. The main limitation is the effect of confounding factors[198], as many drug-medical event pairs that are associated do not correspond to ADRs. The existing methods that signal drug-medical event pairs based on association do not consider the eight other Bradford-Hill factors, but some of these could be integrated by utilising the data available in the LODs. The RORR method has the potential to cover the most Bradford-Hill causality factors, but it is a signal refinement method rather than a signal generating method, as it requires a signal generating method such as the ROR to identify which drug-medical event pairs to apply the regression models on. Therefore, the RROR is limited by any limitations with the signal generating method it incorporates.

2.1.4.5 Summary

The LOD algorithms show promise at becoming an integral part of pharmacovigilance in the future due to the wealth of information they potentially hold [198].

Although numerous methods have been developed for generating ADR signals using LODs, they have not been robustly analysed. Their theoretical foundations would suggest that they are likely to signal many non-ADRs due to the reliance on association. The RORR method, presented to identify confounding, requires initial drug-medical event pair signals to be generated, so it is a signal refinement method. There has been no method that combines signal generating and refinement into one, but such a method could signal drug-medical event pairs more efficiently and obtain a lower rate of signalling non-causal relationships.

2.1.5 Combining Multiple Databases

2.1.5.1 Overview

There has been a recent initiative to integrate multiple electronic healthcare data sources into one. Examples include the Mini-Sentinel [136], that will eventually become Sentinel, a US Congress mandated pharmacovigilance system that contains medical data for more that 125 million Americans [124], the Exploring and Understanding Adverse Drug Reactions (EU-ADR) project [34], a European initiative set up in 2008 that contains data on over 30 million patients and the Observational Medical Outcomes Partnership (OMOP) that has a network of databases containing over 200 million patients. Numerous researchers have expressed the significance of large pharmacovigilance sources in aiding the ability to discover ADRs efficiently [144]. The initiatives may bridge gaps in the current pharmacovigilance, such as lack of knowledge concerning drug safety for minority groups [34].

The OMOP was formed to analyse the methodologies for pharmacovigilance

using longitudinal data. The partnership have developed a common data model that enables the combination of different databases by transforming them into a general format [115; 131]. The OMOP have presented a magnitude of different techniques specifically for signalling ADRs using longitudinal data, including cohort studies [107], disproportionality methods [216], case series methods [71], case control methods [71], case crossover methods [160] and propensity score based methods [159]. To enable an analysis of the methods, an approximate gold standard consisting of 53 'ground truth' drug-medical event pairs (i.e., drug-medical event pairs that are known to be ADRs or non-ADRs) have been identified [141]. The ability of the methods to generate correct signals for these 'ground truths', at their natural threshold, has been investigated [156].

The standardised 'ground truths' only consider a selection of medical events, referred to as Health Outcomes of Interest (HOI) and a small subset of drugs known as Drugs Of Interest (DOI). Tables 2.7-2.8 display the OMOP's proposed HOIs and DOIs. Unfortunately, there are few studies investigating the methods abilities in generating signals when a large number of drug-medical event pairs are studied, however this is more realistic [143].

The OMOP methods tend to be based on standard epidemiological studies that aim to identify associations between drugs and medical events by finding medical event that have a greater incidence after a drug compared to the medical event's estimated background incidence. Many methods have been presented and the seven that have been extensively investigated as described below. The first method, the High–throughput Screening by Indiana University (HSIU), is

[1]READ codes do not exist for the exact medical event, so GI ulcer READ codes are given.
[2]READ codes do not exist for the exact medical event, so mortality due to cardiac or patient died READ codes are given.

Table 2.7: The Heath Outcomes of Interest defined by the OMOP [80] and their corresponding THIN READ codes.

Medical event	THIN READ code
Angioedema	SN51.
Aplastic anemia	D20.., D2011, D201., D2012, D2012, D204., D202., Dyu2.
Acute liver injury	J6000, J6357
Bleeding	J68.., J68z., J68z0, J68z1, J68z2, J68zz
GI ulcer hospitalization [1]	J11.., J110's
Hip fracture	S30.., S30y.
Hospitalization	8H2.., 8H2z., 8H7a., 8Hd.., 8HJ.., 9144.
Acute myocardial infraction	G30.., G30's
Mortality after myocardial infarction [2]	G5751, 22J..
Acute renal failure	K04.., K04y., K04z., Kyu20, K043.

a cohort approach [80]. A cohort study follows a group of patients that have a common attribute or event (such as a drug prescription) and assesses outcome risk factors [178]. The Observation Screening method [155] calculates the screening rate within the drug population (the frequency of a outcome divided by the total risk time) and normalises this by dividing it by an estimate for the background screening rate. This is either the screening rate in a non-risk period (frequency in a pre-exposure period divided by the total pre-exposure time period) or the screening rate in a control group. The third method, the Disproportionality Analysis (DP) [216], identifies associations by comparing the rate that a medical event occurs within the drug population relative to the rate it occurs within some other population, similar to the SRS methods in Chapter 2.1.3.

The Univariate Self-control Case Series (USCCS), based on the method developed in [54], can be considered a cohort based study but where the exposed and non-exposed patients are the same. The approach partitions the cases' timelines

Table 2.8: The Drugs of Interest defined by the OMOP, table from [80].

DOI Drug Name	DOI Description
OMOP ACE Inhibitor	ACE inhibitors: benazepril, captopril, enalapril, fosinopril, lisinopril, quinapril, and ramipril; restricted to oral form
OMOP Amphotericin B	parenteral Amphotericin B
OMOP Antibiotics: erythromycins, sulfonamides, and thetracyclines	Antibiotics: erythromycins, sulfonamides, and tetracyclines; restricted to oral and injectable
OMOP Antiepileptics: carbamazepine, phenytoin	Antiepiletics: carbamasepine, phenytoin: restricted to oral and injectable
OMOP Benzodiazepines	Benzodiazepines: alprazolam, chlordiazepoxide, clonazepam, clorazepate, diazepam, estazolam, flurazepam, halazepam, lorazepam, oxazepam, prazepam, quazepam, temazepam, or triazolam
OMOP Beta blockers	Beta blockers: propanolol, metoprolol, atenolol; restricted to oral form
OMOP Bisphosphonates	Bisphosphonates: alendronate
OMOP Tricyclic antidepressants	Tricyclic antidepressants: restricted to oral and injectable
OMOP Typical antipsychotics	Typical antipsychotics: Chlorpromazine, chlorprothixene, levomepromazine, flupentixol, Fluphenazine decanoate, Fluphenazine enanthate, Fluphenazine hcl, Haloperidol, Haloperidol decanoate, Loxapine hcl, Loxapine succinate, melperon, Mesoridazine, Molindone, Perphenazine, amitriptyline hcl/perphenazine, Pimozide, pipamperone, promazine, Prochlorperazine edisylate, periciazine, Prochlorperazine maleate, Promazine, Propiomazine, Thioridazine, Thiothixene, Trifluoperazine, zuclopenthixol
OMOP Warfarin	Warfarin

into hazard and non-harard periods and compares the incidence in the hazard periods with the incidence in the non-hazard periods [200]. Fixed in time confounding is overcome within the USCCS by using the same patients as the exposed and non-exposed. The Multi-Set Case Control Estimation (MSCCE) method is a case control approach that selects cases based on the occurrence of a specified condition and selects control that do not have the condition and are active over the required observation period (i.e., have events reported before and after the period) [217]. The Bayesian Logistic Regression (BLR) [87] method applies a logistic regression approach using prior knowledge to initialise the coefficients that determine the weight that each covariate has on the final output. The final method is the Information Component Temporal Pattern Discovery (ICTPD), summarised in Chapter 2.1.4.2.

The seven methods only cover the Bradford-Hill association strength consideration and some incorporate filters to cover the temporality. Consistency is indirectly incorporated due to the combination of multiple data sources. Some of the methods, such as the BLR, remove confounding by adjusting for covariates or apply stratification to reduce confounding by age and gender.

The seven OMOP methods described above were applied in the Non-Specific Association (NSA) experiment, whereby the ten DOIs were paired with all possible outcomes and the signals generated by each method at their natural thresholds were determined [80]. The majority of methods have many parameters that determine their performance and the study applied the methods over a range of parameter values to identify the optimal performance. This shows that additional work is required to tune these existing methods depending on the database being used. The performance of the seven OMOP methods during the NSA ex-

Table 2.9: The OMOP methods NSA experiment results.

Method	Optimal Scores over the NSA experiment			
	AUC	MAP	P(10)	FPR
HSIU	0.7342	0.1408	0.42	0.2658
OS	0.7138	0.0942	0.22	0.2862
DP	0.6741	0.0622	0.23	0.3259
USCCS	0.7342	0.1408	0.4200	0.2658
MSCCE	0.603	0.032	0.05	0.397
BLR	0.6329	0.0316	0.03	0.3671
ICTPD	0.6695	0.0591	0.1	0.3305

periment is presented in Table 2.9. It can be observed that all seven methods had False Positive Rates (FPRs) greater than 0.25 and Mean Average precision (MAP) scores less than 0.015. The AUC values ranged from $0.6 - 0.735$, as the AUC corresponds to the probability that an ADR is ranked above a non-ADR (rank 1 being the highest) [20], there is still approximately 30%-40% chance than a non-ADR will be ranked higher than an ADR.

A recent study investigated potential loss from mapping the raw THIN data into the common data model [214]. A few existing methods were applied to both the raw THIN database and the THIN database mapped to the common data model. The results of the study suggested that the existing methods performed equally well on the raw and mapped data when considering the signals generated for the 53 ground truths. However, the study showed that 55% of drug codes and 25% of medical events codes could not be mapped from THIN into the common model [214], and this is likely to have detrimental effects when more than the 53 ground truths are considered. This highlights the important of developing database specific methods, in additional to the common model methods, that can utilise all the data available and present an alternative perspective for ADR

discovery. When improvements in the mapping to the common data model are developed, then any method developed for THIN could also be modified for implementation on the THIN mapped to the common data model (or any other common data model mapped database).

2.1.5.2 Summary

Combining the databases means that it may be possible to generate signals efficiently [3]. However, the combination requires the data to be transformed and normalised and this has the potential to lose information and can negatively impact the efficiency of signalling ADRs. It was demonstrated in [214] that many of the raw THIN data cannot be incorporated into the common data model, motivating the development of methods that are specific to certain databases. Comparisons of existing OMOP methods have shown that they perform moderately on the common data model [156] and there is no optimal method. In addition, the methods had a high false positive rate, even when the number of drug-medical event pairs being investigated is controlled. It is likely that the methods will be further hindered when applied to determine a drug's complete set of side effects as there will be a surplus number of drug-medical event pairs corresponding to non-ADRs.

2.1.6 Pharmacovigilance Summary

Adverse drug reactions are becoming an increasing burden on the NHS [166]. Existing post-marketing surveillance of drugs is limited by underlying issues associated with SRS databases [79]. Many ADRs are only being found years after the drugs are marketed and as a result, many patients suffer serious health issues

that could be avoided with improved ADR knowledge. Rare ADRs that are hard to identify, ADRs corresponding to medical events with a high background rate or less serious ADRs may never be detected by data-mining algorithms applied to SRS databases [173]. As a result there has been a recent demand for improved post-marketing surveillance [129; 190].

One recent solution has been to develop data-mining algorithms for LODs or to combine multiple electronic healthcare databases as a resource for ADR detection. Unfortunately the current methods developed for LODs have a high false positive rate [156] and have not been extensively investigated due to a lack of a complete 'gold standard' [33]. The high false positive rate is probably due to confounding caused by the countless number of possible covariates. Integrating the Bradford-Hill causality considerations into a signalling method is one possible consideration to reduce the negative impact of confounding factors and therefore reduce the number of false positive signals. The Bradford-Hill causality considerations have been used to help distinguish between associations that are causal, and those that are not. As confounding causes the associations that are non-causal, the Bradford-Hill causality considerations must be able to indirectly identify some confounding. The majority of existing methods only cover a few of the Bradford-Hill causality considerations, however, there is potential to extract data from the LODs to enable novel methods that cover more of the criteria. This could then reduce the number of false positives.

The THIN database is a LOD that contain medical data for over 10 million patents, often spanning decades of years per patient. The general benchmark for the THIN database is unknown, as only a few methods have been investigated by considering the signals generated for a small set of 54 'ground truth' drug-medical

event pairs [214]. A specific method to signal ADR using the THIN database may generate novel signals that cannot be generated using the common data model nor the SRS databases. There are inconsistencies in the recording of data into the THIN database [78], but this may be overcome by developing a novel method that takes this into account.

2.2 Pattern Recognition

In the previous part of the literature review the existing pharmacovigilance techniques that tend to signal ADRs by calculating an estimate for the relative risk of each drug-medical event pair were summarised. The medical events with a large estimated relative risk are then signalled, or alternatively 'classified', as potential ADRs. These methods can be considered unsupervised learning algorithms, algorithms that infer hidden structure without being taught [2], as they do not use knowledge of existing ADRs to learn intrinsic differences between ADRs and non-ADRs. Rather, they use a single attribute such as the relative risk estimate to distinguish between ADRs and non-ADRs. The limitation with relying on a single attribute, such as the relative risk, is that confounding can occur and cause many non-ADRs to have a high relative risk estimate. This results in the techniques having high false positive rates and reduces the efficiency in detecting ADRs.

There has been no research to date that extracts attributes for drug-medical event pairs from LODs and then uses known ADRs as a means to learn the unknown ADRs based on their attributes, although in [113] the authors use chemical knowledge and learn from known ADRs. This type of learning is called supervised learning [74]. During the training stage, supervised learning requires attributes that describe each data-point and knowledge of the 'classes' that the data-points belong to. In the context of ADR signalling each drug-medical event pair would represent a data-point and their attributes would correspond to values that could be used to distinguish between ADRs and non-ADRs. Examples of suitable attributes include the risk of the medical event within a defined time period after

taking the drug or the average age of the patients experiencing the medical event after the drug. Labels need to be assigned to each data-point (i.e., each drug-medical event pair) to define their class, for example the pair ciprofloxacin and tendon rupture would be in the class ADR whereas the pair ciprofloxacin and normal menopause are in the class non-ADR. In the pharmacovigilance field this has been unexplored in general due to the uncertainty with knowing what medical events are definitely ADRs or non-ADRs of a drug . If a sufficient number of labelled data-points could be generated then a supervised algorithm could be trained. This would enable classification of any drug-medical event pair whose ADR status is unknown, as an ADR or non-ADR. If suitable attributes were chosen so that it was possible to distinguish between medical events linked to drugs due to confounding factors and true ADRs, then a supervised algorithm could offer significant improvement over existing ADR signalling methods .

In the following section the theory behind supervised learning and the main algorithms applied are described. This is followed by a summary on semi-supervised learning, the technique developed to deal with the situation of having labels that are difficult to generate [30]. Due to the conundrum that applying supervised learning for signalling ADRs imposes, requiring knowledge of ADRs to extract knowledge of ADRs, it may be impossible to generate the required number of labeled data-points and a semi-supervised algorithm may be more appropriate.

2.2.1 Supervised Learning

2.2.1.1 Introduction

Supervised learning is the process of learning from examples to infer the relationship between inputs and outputs. A training set consisting of inputs (also known as attributes) and their corresponding outputs are used to 'supervise' the training of a function that is capable of generalising the mapping between input and output. The trained function can then be used to predict the output of any unseen input coming from the same distribution as the training set inputs. When the outputs are discrete they are referred to as classes or labels and the supervised learning is known as classification. Alternatively, when the output is continuous the supervised learning is known as regression [37]. For example, if the odds ratio (OR) and risk difference (RD) attributes are known for a thousand different drug-medical event pairs and for each pair their class (ADR or non-ADR) is also known, then supervised learning could be applied to partition the attribute space into areas likely to correspond to ADRs and areas likely to correspond to non-ADRs, see Fig 2.9 illustrating the ideal situation where ADRs and non-ADRs are separable in the space determined by the OR and RD.

Formalising the previous statement, the training set A_n is a collection of inputs $\mathbf{x_i} \in X$ and corresponding outputs $y_i \in Y$ pairs,

$$A_n = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_n}, y_n)\} \tag{2.12}$$

Where each pair $(\mathbf{x_i}, y_i)$ are assumed to be independent identically distributed samples from an unknown joint probability distribution P. In general, $X \subset \mathbb{R}^m$

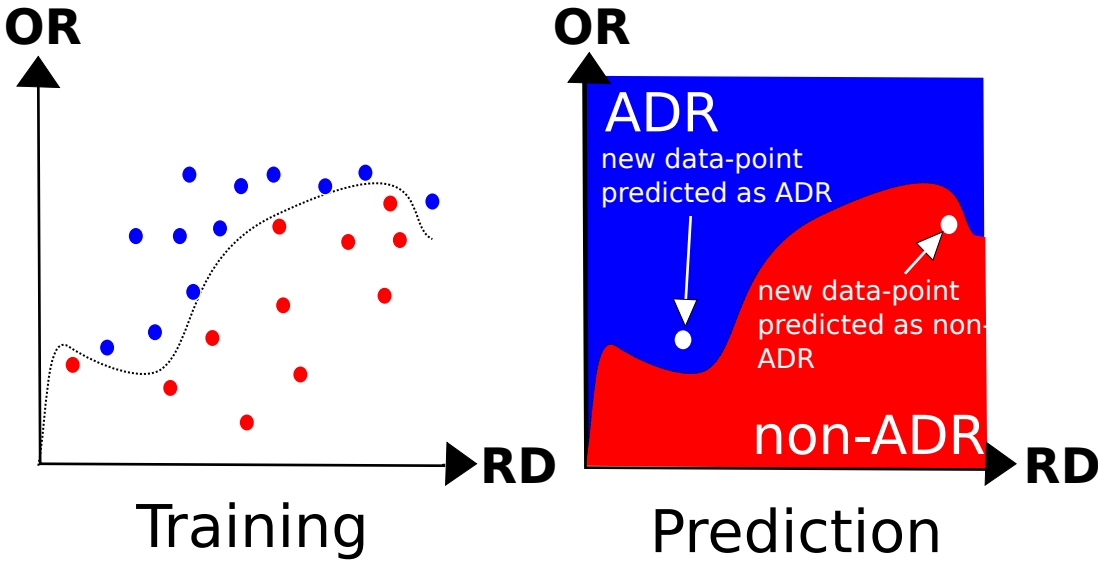


Figure 2.9: Illustration of a classifier partitioning the attribute space. Using the training data (blue dots are labelled as ADR and red as non-ADR) a function is trained to partition the space into ADR sections and non-ADR sections. This can then be used to predict whether a new data-point is an ADR or non-ADR based on where the data point lies in the attribute space.

and $Y \subset \mathbb{R}$ for regression or $Y = \{-1, 1\}$ for binary classification. The task of supervised learning is to find a function $f : X \to Y$, where $f \in H$ (a class of functions).

As the training data is considered to consist of n independently identically distributed samples from an unknown joint probability distribution $P(\mathbf{x}, y)$, then the task of supervised learning is to develop a function f that models the dependency within the joint distribution. There are two different approaches for producing machine learning models, the discriminative model and the generative model. The discriminative model aims to determine the conditional distribution of the class label given the input, $P(y|\mathbf{x})$, by using a parametric model and determining the model's parameter values with the aid of the training set [102]. The generative method calculates the joint probability distribution, $P(\mathbf{x}, y)$, and makes use of this distribution to predict the conditional distribution [102]. In general, if the training set is sufficiently large (depending on the complexity of the model), discriminative models have been shown to perform better [92], however, generative models have the advantage of being able to incorporate unlabelled data [102]. This is advantageous when generating labels becomes costly.

To find the optimal function $f \in H$ for mapping the inputs to their outputs it is necessary to evaluate each functions performance. This is calculated by a non-negative loss function,$L : Y \times Y \to \mathbb{R}^+$, that determines a measure of error between the predicted output $f(\mathbf{x_i})$ and the real output y_i. Various loss functions have been proposed, examples for binary classification [150] include,

- Square Loss: $L(f(\mathbf{x}), y) = (1 - f(\mathbf{x})y)^2$
- Hinge Loss: $L(f(\mathbf{x}), y) = |1 - f(\mathbf{x})y)|_+$

- Logistic Loss: $L(f(\mathbf{x}), y) = (ln2)^{-1} ln(1 + e^{-f(\mathbf{x})y})$

The choice of the loss function that is implemented should be chosen based on the specific classification problem [4]. The integral of the model's loss function over the joint probability distribution gives the generalisation error, or risk,

$$R(f) = \int L(f(\mathbf{x}), y) dP(\mathbf{x}, y) \tag{2.13}$$

The Bayes estimator, g^*, is the function that minimises the risk,

$$R(g^*) = \inf_f R(f) \tag{2.14}$$

The goal of a discriminative learning algorithm is to find the function within a class of possible functions, $f^* \in H$, that minimises the risk, $f^* = \arg\min_{f \in H} R(f)$. Unfortunately it is often the case that the Bayes estimator does not belong to the class of possible functions. Methods that aim to determine the function that minimises the risk include empirical risk minimisation [123], structural risk minimisation [187], regularisation [18] and normalised regularisation [18].

Empirical risk minimisation is a simple and generally efficient means to determine a suitable function. The empirical risk measures the difference between the predicted output values and the true output values by calculating the average of the loss function over each data-point in the training set,

$$R_{emp}(f, A_n) = \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{x_i}), y_i) \tag{2.15}$$

The empirical risk minimisation method then identifies the function f from a

model of possible functions H that minimises the empirical risk,

$$f^* = \arg\min_{f \in H} R_{emp}(f)$$

It is clear that the choice of model that determines the possible functions has a direct impact on the results returned by the empirical risk minimisation method.

The idea behind the structural risk minimisation is to pick a sequence of models, $\{H_s | s \in \mathbb{N}\}$, than increase in size and find the argument that minimises a trade off between the empirical risk and a penalty that penalises large models (models with a large capacity),

$$f^* = \arg\min_{f \in H_s, s \in \mathbb{N}} R_{emp}(f) + pen(s, n) \tag{2.16}$$

where n is the size of the training data. As the empirical risk only estimates the actual risk, it is of interest to find bounds on the difference between the actual and empirical risk, as this gives an indication into the predictive suitability of any functions that are determined using a supervised learning model. Extensive analysis by [187] managed to show that the actual risk is bounded by the empirical risk and an additional term than corresponds to the complexity of the model. With probability $1 - \eta$ the following holds,

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(log(2n/h) - log(\eta/4))}{n}} \tag{2.17}$$

where h is the VC dimension of the class of functions H, this is a measure of their complexity. The VC dimension of a class of functions is the maximum number of points that can be separated in every possible way by those functions over a

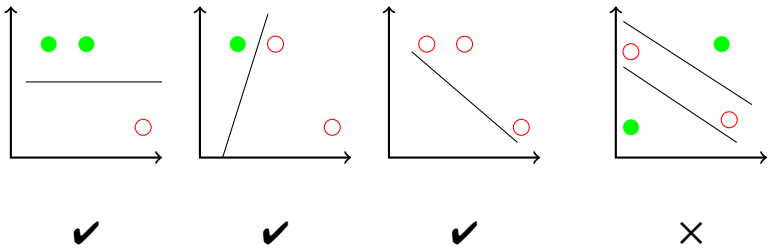


Figure 2.10: Illustration of the maximum number of points separable in every possible way by a linear classifier.

defined space [189]. A visual example demonstrating that a linear classifier has a VC dimension of 3 can be seen in Figure 2.10. It can be seen that 3 non-collinear points can be separated by a line in every possible way, but this is not the case for 4 points, as the far right graph shows two lines are required.

In general the bound can be represented as,

$$\text{Test error} \leq \text{Training error} + \text{Complexity of set of models} \tag{2.18}$$

Training a highly complex model may lead to overfitting, where the training error is minimised but the model is not generalised and performs poorly on the testing data. On the other hand, a less complex model is likely to have a high training error. Therefore, the perfect model determines a function that has a low training error but is also as simple as possible.

The complexity of the model depends on H, the class of functions, and this is determined by the classifier being applied. The most widely applied classifiers are the Decision Tree [81], Naive Bayes [101], Logistic Regression [84], Support Vector Machine [39] and K-Nearest Neighbours (KNN) [56]. Each of these have different model assumptions and are briefly summarised in the following section.

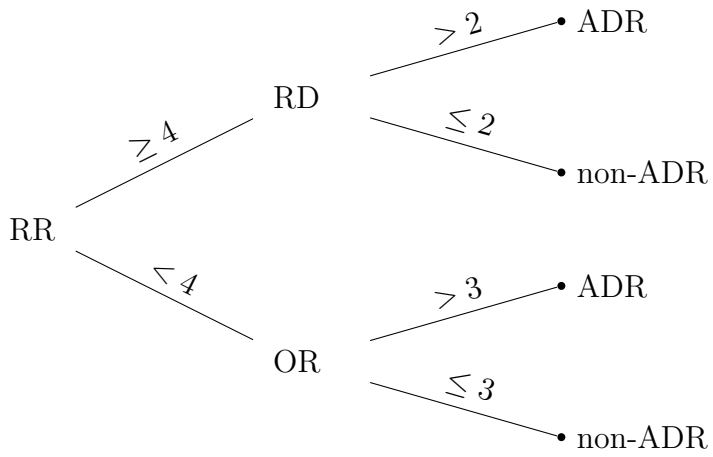


Figure 2.11: Example of a decision tree to classify drug-medical event pairs as ADRs or non-ADRs.

2.2.1.2 Classifiers

Decision Tree

The decision tree classifier is a directed tree that recursively partitions the attribute space into sub-spaces. An illustration of a hypothetical decision tree can be seen in Figure 2.11. A decision tree is non-parametric [116], self explanatory [116] and has the advantage of being unaffected by heterogeneous data or different features that have varied ranges. This means that the data does not need to be extensively processed before applying the classifier. Unfortunately, it has been described as 'greedy' as noise or irrelevant attributes in the training set can greatly impede its performance [139].

The decision tree can be constructed with a bottom-up [95] or top-down [149] approach. Generally speaking, the algorithm uses a splitting measure to calculate how well an available partitioning of the space separates the classes. During each iteration in the top-down approach, the optimal partitioning is applied to the current subspace, or the subspace stops being partitioned when the splitting

measures shows there is no possible partition that can lead to a sufficient gain or the stopping criterion is satisfied. In general, the splitting criteria is only based on a single attribute during each iteration. This is known as univariate splitting [116] and the measures are often based on impurity based criteria such as the Gini index [21] or information gain [138].

The information gain takes its origin from information theory and measures the change in the entropy value that is caused by partitioning the space. The entropy value corresponds to the uncertainty within a set. Considering the binary classification problem where there are two class, let p_1 and p_2 represent the proportion of the data-points within the set S that are in class 1 or -1 respectively, then the entropy is,

$$E(S) = -\sum_{i=1}^{2} p_i log_2 p_i \tag{2.19}$$

If the data-points in a set are all from one class, without loss of generality, assume they are from class 1, then $p_1 = 1$ ($log_2 p_1 = 0$) and $p_2 = 0$ so $E = 0$, the lowest possible value. If the data-points in a set are spread equally between the two classes, $p_1 = p_2 = 0.5$, then the entropy is the highest possible value $E = 1$. It is clear that choosing a partitioning with the highest information gain minimises the entropy and leads to a final partitioning of the space into numerous subspaces that are dense in a single class. The main limitations of using information gain as the splitting measure for a decision tree classifier is that there is a bias towards partitioning based on attributes with large ranges [201] that can lead to over-fitting and it is common for the space to be fragmented into a surplus number of small subspaces.

The Gini index is another splitting measure frequently implemented. The

Gini index is calculated for a set S by,

$$Gini(S) = 1 - \sum_{i=1}^{2} p_i^2 \tag{2.20}$$

The Gini index is minimised when the majority of the data-points within set S belong to one class. In this case, one of the p_i values will be close to one and the others will be small. The square term in the Gini index calculation puts more emphasis on larger values. Squaring the p_i value close to one has little effect, whereas the closer a p_i value is to zero, the more it becomes reduced when squared. So a set S containing data-points spread between different classes will have a small value for $\sum_{i=1}^{2} p_i^2$ and therefore a Gini index close to 1.

The average Gini index is the weighted average of the Gini index based on partitioning the set S into subsets S_i using the values of a single attribute A, where $|S|$ corresponds to the number of elements in the set S,

$$Gini(S, A) = \sum_{i} \frac{|S_i|}{|S|} Gini(S_i) \tag{2.21}$$

The decision tree is generated by finding the partitions that minimise the average Gini index. Research comparing the different univariate splitting measures has often concluded that the choice has little effect on the decision tree as there does not appear to be an overall superior measure [116].

Naive Bayes

The Naive Bayes classifier uses the training set to determine the distribution of the class label, $P(Y)$, and the conditional distribution of the input attributes

given the class label, $P(X_i|Y)$ for $i \in [1, n]$, and then use these combined with Bayes rules and a conditional independence assumption to find the most probable class for any future inputs. The conditional independence assumption is used to simplify the number of parameters required by the model and enables an efficient calculation of the distribution $P(Y|X)$.

Consider three random variables, X, Y and Z. It is defined that X is conditionally independent of Y given Z if, $P(X|Y, Z) = P(X|Z)$. Assuming that the input features are conditionally independent given the class label, then,

$$\begin{aligned} P(X_1, X_2, ..., X_n|Y) &= P(X_1|X_2, ..., X_n, Y)P(X_2|X_3, ..., X_N, Y)...P(X_n|Y) \\ &= P(X_1|Y)P(X_2|Y)...P(X_n|Y) \\ &= \prod_{i=1}^{n} P(X_i|Y) \end{aligned}$$

Using Bayes rule,

$$P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k)P(X_1, ..., X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1, ..., X_n|Y = y_j)} \tag{2.22}$$

and using the conditional independence the expression for the conditional probability of the class label given the input data is,

$$P(Y = y_k|X_1, ..., X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)} \tag{2.23}$$

As the denominator in equation (2.23) is independent of the choice of class label, it can be ignored. The classifier simply determines the most probable class label

for an input by,

$$Y = \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i|Y = y_k) \tag{2.24}$$

The classifier, although limited by its unrealistic assumption of conditional independence, has performed well for some real life classification problems [47]. In [146] the authors state that a known limitation of the Naive Bayes classifier is that it does not perform optimally when the classes are non-linearly separable.

Logistic Regression

Logistic regression is a discriminative model so it assumes a distribution for $P(Y|X)$ and uses the training data to determine the parameter values. Logistic regression is applied when the classification is binary and does not required the inputs to be normally distributed, have equal variance within each class nor be linearly related [27]. The main disadvantage with the classifier is that it uses maximum likelihood to determine the parameter values and this requires larger training sizes than for linear regression. It is suggested that a minimum of 50 cases per predictor are used [27].

The logistic regression model is based on the assumption that the log odds of a data-point belonging to a class given its n attributes can be expressed as a linear combination of the data-points attributes. Under this assumption the log odds is expressed as,

$$ln(P(Y|\mathbf{X})/(1 - P(Y|\mathbf{X}))) = w_0 + \sum_{i=1}^{n} w_i X_i$$

by taking the exponential and re-arranging, the conditional model used by logistic regression for the two class problem is,

$$P(Y=0|X) = \frac{exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$\begin{aligned} P(Y=1|X) &= 1 - P(Y=0|X) \\ &= \frac{1}{1 + exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \end{aligned}$$

As the classifier assigns the class for an input $\mathbf{x}$ based on $\arg\max_{y_k} P(Y = y_k|X = \mathbf{x})$, it is clear that class 0 is assigned when $1 < exp(w_0 + \sum_i w_i X_i)$ (or equivalently $0 < w_0 + \sum_i w_i X_i$) and class 1 is assigned otherwise.

Support Vector Machine

The Support vector machine (SVM) classifier is a parametric model that aims to find the hyperplane that separates the classes while maximising the distance between the data-points and the hyperplane, see Figure 2.12. For the two class problem, the SVM works by finding two parallel hyperplanes such that they separate the two classes and there are no points between the two hyperplanes. The equations of two hyperplanes are $\mathbf{w}.\mathbf{x_i} + b = 1$ and $\mathbf{w}.\mathbf{x_i} + b = -1$. The bit in-between the hyperplanes is referred to as the 'margin'. This is what needs to be maximised to ensure the classes are separated as much as possible. As the distances between the two hyperplanes is $2/||w||$, by minimising $||w||$ we can find the maximum separation between the classes. Previously in Chapter 2.2.1.1 it was shown that the actual risk of a classifier is bounded by the empirical risk and a term that depends on the capacity/complexity of the set of decision functions

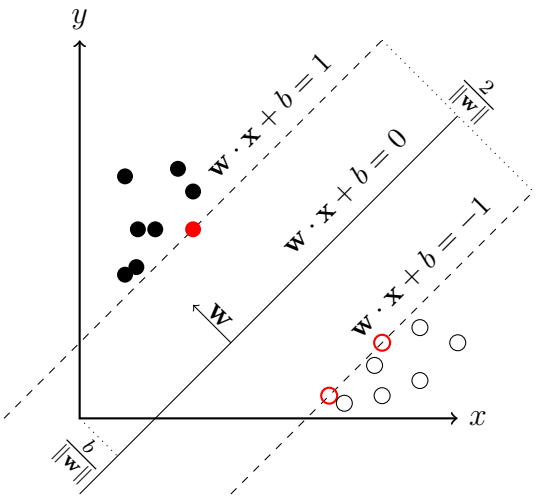


Figure 2.12: Illustration of the support vector machine classifier. The hyperplane separating the classes is positioned such that the distance between the hyperplane and the closest data points from either class is maximised.

defined by the classifier. The decision functions used by the SVM classifier are the hyperplanes $\mathbf{w}.\mathbf{x} + b$. It has been proven that for the set of hyperplanes $(\mathbf{w}.\mathbf{x}) = 0$ such that $\min_i |\mathbf{w}.\mathbf{x}_i| = 1$ for $\mathbf{x_i} \in X$ the set of decision functions $f_w(\mathbf{x}) : X \to \{-1, 0, 1\}; f_w(\mathbf{x}) = sgn(\mathbf{w}.\mathbf{x})$ satisfying $||\mathbf{w}|| < A$ has a bounded VC dimension [188],

$$h \leq R^2 A^2 \tag{2.25}$$

where R is the radius of the smallest ball centred at the origin that covers the set X. This defines an upper bound on the capacity/complexity of the SVM classifier. In the separable case this motivates finding the parameters $\mathbf{w}$ and b such that $||\mathbf{w}||^2$ is minimised and,

$$\begin{aligned} (\mathbf{w}.\mathbf{x_i} + b) &\geq 1 \quad \text{if } y_i = 1; \\ (\mathbf{w}.\mathbf{x_i} + b) &\leq -1 \quad \text{if } y_i = -1. \end{aligned} \tag{2.26}$$

as this results in a zero empirical risk and minimises the capacity of the model. When the data is non-separable (i.e. there is no hyperpane that can cleanly split the classes), slack variables are introduced to enable the misclassification of some data-points. The aim is to maximise the margin while minimising the degree of misclassification. The optimisation problem becomes minimise:

$$||\mathbf{w}||^2 + C\sum_{i=1}^{m}\sigma_i$$

subject to:

$$y_i(\mathbf{w}.\mathbf{x_i} + b) \geq 1 - \sigma_i, \sigma_i \geq 0$$

An addition to the optimisation problem includes incorporating kernel functions that map the data-points into a space where they are separable [125].

K-Nearest Neighbour

The K-nearest neighbour (KNN) classifier is non-parametric, this means it does not assume the data come from a specific distribution. The classifier is described as a lazy algorithm as it does not use the training data to generalise (generate a probabilistic distribution) [199], this makes the training state highly efficient, but can cause the testing step to become costly.

The classifier requires the data to come from a metric space, but the measure of distance can be any suitable metric. The classifier works by taking the majority vote of the k nearest neighbours, where distance is determined by the defined metric. If the set $N_k(\mathbf{x})$ is the set of indices corresponding to the K nearest

neighbours of $\mathbf{x}$, then,

$$f_{KNN}(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i \in N_k(\mathbf{x})} y_i \geq 0; \\ -1 & \text{if } \sum_{i \in N_k(\mathbf{x})} y_i < 0. \end{cases}$$

For example, if $k = 7$, and for an input $\mathbf{x}$ three of its neighbours are class -1 and four are class 1 then the input would be assigned the class 1. The algorithm can be modified to use the distance of the neighbours as weights so that closer neighbours have more influence [49].

2.2.1.3 Ensemble Methods

An ensemble classifier considers the outputs from multiple trained classifier to determine the class of a data-point [130]. In general, the method combines multiple diverse 'weak learners' to produce a 'strong learner'. The motivation behind an ensemble classifier is to reduce the bias that can occur when considering single classifiers and to reduce the variance than can occur due to the choice of data used during training [23].

There exists a magnitude of options for generating diverse classifiers including building models from different samples of the data [22], using different models [51] or building models that use different subsets of attributes [82]. There are also different ways to combine the predictions from the classifiers, such as determining the class by voting that returns the modal class or weighted voting that incorporates the confidence of the classifiers or error estimations as weights to produce a weighted sum of the votes. Another method, known as stacking, is to use the outputs of the classifiers as inputs into a new meta classifier that does

the final classification [177].

The most widely implemented ensemble methods that use sampling of the training data set are known as bagging [22] and boosting [60]. Bagging involves iteratively generating classifiers that are built on different training sets and returning the class with the highest number of votes based on these classifiers. The different training subsets are produced by drawing with replacement from the whole training set. Bagging has a statistical basis and can be considered similar to averaging as it reduces the classifier's variance [23; 130]. The advantage of bagging is that is it resistant to noise, however, experiments have shown that with a little noise present it is not as accurate as other methods such as boosting [46].

Boosting has its foundations in learning theory and the general aim is to produce a sequence of classifiers that are used to generate a weighted vote for the overall class. The misclassifications of the previous classifiers in the sequence have an influence on the weights assigned during classification in the later sequence classifiers. The most widely used boosting classifier is the AdaBoost classifier developed by Freund and Schapire [61] that generates a sequence of simple classifiers ($h_m \in H$, where H is a class of simple classifiers) and weights ($\lambda_m \in \mathbb{R}$) by giving more importance to data-points that were misclassified by the simple classifiers earlier in the sequence. The final classification makes use of the weighted majority vote $\text{sgn}(\sum_{m=1}^{M} \lambda_m h_m(x))$. Considering $(X_i, Y_i), i \in [1, n]$ to be i.i.d. samples where $Y_i \in \{-1, 1\}$ and $X_i \in x$ then the sequence is determined by,

0. Let $c_1 = c_2 = ... = c_n = 1$, and set $m = 1$.

1. Find $h_m = \arg\max_{h \in H} \sum_{i=1}^{n} c_i h(X_i) Y_i$. Set

$$\lambda_m = \frac{1}{2} log(\frac{\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i h_m(X_i) Y_i}{\sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i h_m(X_i) Y_i}) = \frac{1}{2} log(\frac{\sum_{h_m(X_i)=Y_i} c_i}{\sum_{h_m(X_i)\neq Y_i} c_i}) \quad (2.27)$$

2. Set $c_i \leftarrow c_i exp(-\lambda_m h_m(X_i) Y_i)$, and $m \leftarrow m+1$, If $m \leq M$, return to step 1.

In step 1 the algorithm finds the simple classifier in H that has the smallest weighted misclassification and then calculates the corresponding lambda based on the ratio of correct classifications to misclassifications. The weights that determine the importance of correctly classifying each datapoint are then updated in step 2. If the simple classifier misclassified datapoint i then $-\lambda_m h_m(X_i) Y_i$ will be positive and therefore the weight given to it will increase, alternatively if the classifier was correct the weight will decrease. Boosting has been shown to often work well but it has been hypothesised that results by boosting may be impeded when there is noise present in the training set [46; 59].

The random forest is a non-parametric ensemble classifier that produces a 'forest' containing multiple decision trees and determines the class based on majority voting whereby each tree in the forest is given one vote [24]. Each decision tree is built on a different random sample of the training set, where sampling is done with replacement.

Let the training set $D_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_n, Y_n)\}$ consists of n i.i.d. pairs of random variables sampled from the joint distribution $(\mathbf{X}, Y)$ where $X = \mathbb{R}$ and $Y = \{0, 1\}$. We represent the marginal distribution of X by $\mu(x) = P\{X = x\}$ and the posteriori probability by $\eta(x) = P\{Y = 1 | X = x\}$. The probability

of a classifier g_n misclassifying is,

$$L(g_n) = P\{g_n(X, D_n) \neq Y\}$$

It has been shown that the Bayes classifier, $g^*(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}$, minimises the probability of error [45] and this probability of error for the Bayes classifier $L(g^*)$ is referred to as the Bayes risk. A sequence of classifiers $(\{g_n\})$ is consistent for the distribution $(\mathbf{X}, Y)$ if $\forall \epsilon > 0 \; \exists N \in \mathbb{N}$ s.t. $\forall n \geq N \; |L(g^*) - L(g_n)| < \epsilon$.

A randomised classifier $g_n(X, \theta, D_n)$ uses a random variable θ to determine its prediction, where θ takes its values from some measurable space. The probability of error for the randomised classifier can be calculated as,

$$L(g_n) = P\{g_n(X, \theta, D_n) \neq Y | D_n\}$$

Given m identically distributed draws from the random variable θ, $\theta^m = (\theta_1, ..., \theta_m)$ where each of the $\theta_i s$ are considered independent conditionally on X, Y and D_n, the random forest classifier is constructed such that it takes the majority vote of m decision trees by,

$$g_n^{(m)}(x, \theta^m, D_n) = \begin{cases} 1 & \text{if } \frac{1}{m} \sum_{j=1}^{m} g_n(X, \theta_j, D_n) \geq \frac{1}{2} \\ 0 & \text{else} \end{cases} \tag{2.28}$$

In [13], they prove that if the sequence of random classifiers is consistent then so is the voting classifier. This result implies that if the sequence of random decision trees generated by the random forest is consistent then the probability of error of the random forest tends to the Bayes risk as the number of random

trees increases. One example of the randomisation procedure used to generate the random classifiers by the random forest is to use bagging. In this case each decision trees is built on a random sample of the training data. Another common method is to randomly sample from the attributes available in the training data and train each tree on a difference set of attributes. Some random forest classifiers incorporate the randomness by generating decision trees that interactively pick a random attribute to partition the attribute space until each partition only contains a single data-point from the training set, and then the class returned for a new data-point is the class of the training data-point corresponding the the subspace that the new data-point is in [24]. This method has been shown to have similarities with the nearest neighbour classifier [110].

2.2.1.4 Supervised Learning Summary

In this section the statistical learning theory undermining supervised learning was summarised and the main supervised classifiers currently implemented were presented. It is clear that given sufficient historical data it is possible to learn underlying patterns within the data that can be used to form future predictions. As the THIN database contains a large quantity of historical data, supervised learning techniques can be applied with the aim of inferring medical information that can help improve current healthcare. In the later parts of this section the ensemble classifiers, that make use of multiple classifiers with the aim of improving the classifying accuracy on average, were discussed. In particular, the focus was aimed towards the random forest as this classifier can be applied to heterogeneous data and by incorporating bagging it can be more resilience to noise. These are the two key issues associated with the THIN database, suggesting the random

forest may have excellent performance when applied to classify ADRs using the THIN database.

The majority of existing algorithms for signalling ADRs using electronic healthcare databases are unsupervised as they do not include known ADR labels when detecting patterns and instead find general structures of interest within the data. The reason few supervised algorithms exist is due to the lack of known ADRs preventing the ability to have sufficient quantities of labelled data. However, if these labels can be discovered then a supervised algorithm, with appropriate attributes, may significantly outperform its unsupervised counterpart. Due to clinical trials and knowledge gained over the time that a drug is actively prescribed, some ADRs are definitively known and could be used as labels. If there are some labels but not enough, then an alternative method would be to apply semi-supervised learning. Semi-supervised learning is a mixture of supervised and unsupervised learning techniques. In involves the inclusion of unlabelled data-points into the training stage of an algorithm when there is a small number of labelled data-points [30]. It is often observed that including unlabelled data-points during training can lead to an improvement in performance of the algorithm [30]. This is discussed further in the next section.

2.2.2 Semi-Supervised Learning

2.2.2.1 Introduction

Supervised classification was previously introduced, where the aim is to find a function that approximates the joint distribution between the random variables $\mathbf{X}$ and Y when given n random i.i.d. samples. Unfortunately, it is not always

possible to observe both $\mathbf{X}$ and Y together a sufficient number of times as the label Y can be scarce or costly to determine [30]. When the number of labelled samples in the training set is low, any classifier trained on the data is likely to perform poorly [140].

When the number of labelled data-points are scarce but unlabelled data-points are readily observable, under certain assumptions, knowledge of the marginal distribution can result in an improvement in the function that approximates the joint distribution. Semi-supervised learning algorithms make use of unlabelled data-points to learn the marginal distribution and incorporate this in addition to the labelled data-points when inferring the joint distribution. Formally, given both labelled ($\{(\mathbf{X}_i, Y)\}_{i=1}^{l}$) and unlabelled ($\{\mathbf{X}_i\}_{i=l+1}^{l+u}$) data the aim of the supervised learner is to infer the joint probability distribution $P(\mathbf{X}, Y)$ where the labelled data-points are i.i.d. samples from the joint distribution and the unlabelled data-points are i.i.d. from the marginal distribution $P(\mathbf{X})$. In general there are more unlabelled data-points, $l << u$

In the remainder of this introduction, the main semi-supervised techniques are summarised and the limitations associated to the assumptions they make to enable the incorporation of unlabelled data are discussed.

Self-training Algorithm

The self-training algorithm [40] trains a classifier on the labelled data and then applies the trained classifier on the unlabelled data to predict their class. The algorithm then assumes that some of the predicted classes of the unlabelled data-points are true and moves these from the unlabelled dataset into the labelled dataset. The algorithm continues until the unlabelled dataset is empty. Gener-

ally the algorithm considers the model predictions for the unlabelled data-points with the greatest prediction confidences to be true, however, this is not always the case when the classes are non-separable [215]. Consequently, the self-training algorithm can perform poorly when the classes are non-separable. Early mistakes can have huge impacts as misclassifications will be incorporated into the training of the classifiers in future iterations, potentially leading to further misclassifications.

Probability Generating Models

The aim of each classifier is to identify the most probable class given the input, $\arg\max_Y p(Y|\mathbf{X})$, and this can be determined using a generative model. The generative model makes use of Bayes rule to show that $\arg\max_Y p(Y|\mathbf{X}) = \arg\max_Y p(\mathbf{X}|Y)p(Y)$ and this implies that the class can be determined when the conditional distribution $p(\mathbf{X}|Y)$ and marginal distribution $p(Y)$ are known. If the conditional distribution and marginal distributions are assumed to come from a specified model then given the training data D, the most likely parameter value θ is,

$$\hat{\theta} = \arg\max_{\theta} p(D|\theta) = \arg\max_{\theta} logp(D|\theta) \tag{2.29}$$

and

$$\begin{aligned} logp(D|\theta) &= log(\prod_{i=1}^{l} p(\mathbf{X}_i, Y_i|\theta) \prod_{l+1}^{l+u} p(\mathbf{X}_i|\theta)) \\ &= \sum_{i=1}^{l} logp(Y_i|\theta)p(\mathbf{X}_i|Y_i,\theta) + \sum_{l+1}^{l+u} logp(\mathbf{X}_i|\theta) \end{aligned}$$

The Expectation Maximisation (EM) algorithm [120] can find the value of θ that

locally maximises $p(D|\theta)$. The limitations with the probability generating models are that the probabilistic model needs to be defined and an incorrect model will lead to inaccurate results [215]. It can be difficult to determine the conditional distribution if the number of labelled data-points is small [215]. This technique is more appropriate if there is additional knowledge about the data (e.g., the distribution the data come from is known).

Co-training

The semi-supervised method of co-training [16] is when two different classifiers are trained, in a similar style to self-training except they learn from each other and unlabelled data is iteratively added to each classifier's labeled data based on the predictions of the other classifier. The process of co-training at learning speed k is,

1. Initially let the training sample be $L_1 = L_2 = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_l, Y_l)\}$.

2. Repeat until unlabelled data is used up:

1. Training a view-1 classifier $f^{(1)}$ from L_1 and a view-2 classifier $f^{(2)}$ from L_2.
2. Classify the remaining unlabelled data with $f^{(1)}$ and $f^{(2)}$ separately.
3. Add $f^{(1)}$'s top k most-confident predictions ($\mathbf{X}$,$f^{(1)}(\mathbf{X})$) to L_2.
   Add $f^{(2)}$'s top k most-confident predictions ($\mathbf{X}$, $f^{(2)}(\mathbf{X})$) to L_1.
   Remove these from the unlabelled data.

In effect, the algorithm forces the two classifiers, using different views, to agree on the prediction of the unlabelled data, and the chance of overfitting is reduced.

The co-training algorithm assumes that the data can be partitioned into two different views but how this is done is not always obvious. One limitation of this algorithm is that it requires the two views to be conditionally independent given the class [41],

$$P(\mathbf{X}^{(1)}|Y, \mathbf{X}^{(2)}) = P(\mathbf{X}^{(1)}|Y)$$
$$P(\mathbf{X}^{(2)}|Y, \mathbf{X}^{(1)}) = P(\mathbf{X}^{(2)}|Y)$$

Although in [6] the authors argue that the conditional independance can be relaxed. They suggested that co-training can be applied as long as the two views are not highly correlated. However, many situations are likely to violate this assumption. For example, in the context of classifying a drug-medical event pair as a side effect, if one view uses the knowledge of when the drug occurs relative to the medical event and the other view uses association strength, these views are likely to be highly correlated. If the drug is only observed before the medical event occurs and not after then this will probably mean there is also a strong association between the drug and medical event.

In Chapter 2.2.1.1, the error of a classifier was shown to be bounded by the training error and the term that corresponds to the complexity of the model. It is known that a complex model that minimises the training error may not generalise well to unseen data as it may over-fit. Co-training aims to reduce the complexity of a model by restricting the function space, and therefore reduces the error [215].

2.2.2.2 Semi-Supervised Clustering

When there are no labelled data available, unsupervised techniques such as clustering are applied to find intrinsic patterns within the data [88] without learning from labelled data. Examples include the k-means clustering that initially assigns each data-point into a random cluster and then iteratively moves each data-point into the cluster that is closest [73]. The distance between the data-point and each cluster is based on the cluster centre, the average of the data-points within that cluster. Recent semi-supervised techniques have involved the incorporated of a small number of labelled data to bias the clustering [7]. For example, in [7] the authors use the labelled data to determine the initial centres in the k-means clustering algorithms and fix the labelled data-points into one cluster. Alternative approaches to improving clustering with additional knowledge has involved using must or cannot be in the same cluster constraints [191] or interactive clustering [32], where the semi-supervised clustering algorithm adapts based on feedback.

If the labels are given, the seed-constrained K-means clustering algorithms, developed in [7] improves the unsupervised k-means algorithm by using the labels data to determine the initial cluster centres and then applies the k-means algorithm while fixing the labelled data to their known cluster. The set of data-points input into the seed-constrained K-means algorithm is the set $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$, the value of K input is k (the maximum number of classes in the labelled data) and the initial seeds are $S_l = \{\mathbf{x_i} : \mathbf{x_i}$ is labelled as class l$\}$. The seed-constrained k-means algorithm is described in Algorithm 1.

As there will be labels rather than constrains for the ADR detection problem, the seed-constrained K-means algorithm presents a simple and efficient solution

Input : Set of data-points $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$, $\mathbf{x_i} \in \mathbb{R}^d$, number of clusters K, the set $S = \cup_{l=1}^{K} S_l$ of initial seeds.

Output: Disjoint K partitioning $\{X_l\}_{l=1}^{k}$ of X such that the KMeans objective function is optimised.

Initialization: $\boldsymbol{\mu}_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{\mathbf{x} \in S_h} \mathbf{x}$, for $h = 1, ..., K$; $t \leftarrow 0$

repeat

For $\mathbf{x} \in S$, if $\mathbf{x} \in S_h$ assign $\mathbf{x}$ to the cluster h (i.e., set X_h^{t+1}). For $\mathbf{x} \notin S$, assign $\mathbf{x}$ to the cluster h^* (i.e., set $X_{h^*}^{t+1}$), for $h^* = \underset{h}{\operatorname{argmin}} \|\mathbf{x} - \boldsymbol{\mu}_h^{(t)}\|^2$

$\boldsymbol{\mu}_h^{(t+1)} \leftarrow \frac{1}{|X_h^{(t+1)}|} \sum_{\mathbf{x} \in X_h^{(t+1)}} x$

$t \leftarrow (t+1)$

until *convergence*;

Algorithm 1: The seed-constrained K-means algorithm developed in [7]

if the number of labelled data are low. A common problem with clustering is that the measure of distance this is most suitable for a given problem is generally unknown [210]. The Euclidean distance metric is the standard one implemented, but this treats each attribute equally and assumes the attributes are independent [208]. For many clustering problems these assumptions are unrealistic. This has prompted researchers to develop methods that use the limited number of labelled data available for semi-supervised learning to learn the optimal metric space. By learning the suitable metric space, clustering techniques can be improved [208].

2.2.2.3 Metric Learning

An area of recent research is using additional knowledge to determine the optimal metric, see [99] for a summary. As clustering looks for closely connected communities within the data, the measure of 'closeness' will impact the results, and the standard Euclidean distance may not be most suitable [208]. In [211] the authors proposed learning the metric prior to clustering, whereas in [14] the metric learning is embedded into the clustering and gets applied during each iteration.

In [211], the authors proposed a metric learning algorithm that uses knowledge of constraints (i.e. labelled data-points that are in the same cluster as must-link and data-points that are in different clusters as cannot-link) to learn a mapping from the original attribute space into a new space that maximises the distance between data-points in different clusters while adding a constraint to the maximum distance that data-points in the same cluster can be apart. The algorithm applies eigenvalue optimisation and is highly efficient.

The known constraints are used to determine S, representing the set of all index pairs for data-points that are similar (e.g., $(1,3) \in S$ means that data-point 1 and data-point 3 are known to be in the same cluster), and D, representing the set of index pairs for data-points that are different (e.g., $(1,5) \in S$ means that data-point 1 and data-point 5 are known to be in different clusters). The inner product of two $d \times n$ real valued matrices, $A, B \in \mathbb{R}^{d \times n}$, is denoted by $\langle A, B \rangle := \mathbf{Tr}(A^T B)$, where $\mathbf{Tr}$(A) means the trace of the matrix A and the cone of positive semidefinite matrices is denoted by S_+^d.

Given a pair of data-points $\mathbf{x_i}, \mathbf{x_j} \in X_L^{D_i}$, the matrix $X_{ij} = (\mathbf{x_i} - \mathbf{x_j})(\mathbf{x_i} - \mathbf{x_j})^T$. If $\tau = (i,j)$ is an index pair, then $X_\tau \equiv X_{ij}$. The matrix X_S is defined by $X_S = \sum_{(i,j)\in S} X_{ij}$ and $\widetilde{X}_\tau = X_S^{-1/2} X_\tau X_S^{-1/2}$. The authors calculated that $\nabla f_\mu(S_t^\mu) = \frac{\sum_{\tau \in D} e^{-\langle \widetilde{X}_\tau, S \rangle / \mu} \widetilde{X}_\tau}{\sum_{\tau \in D} e^{-\langle \widetilde{X}_\tau, S \rangle / \mu}}$. The metric learning process, that uses these matrices, is presented in Algorithm 2.

The must-link and cannot-link constrains can be determined when some labelled drug-medical event pairs are known. The must-link pairs are all combination consisting of any two of the known ADRs pairs or all combinations consisting of any two of the known non-ADRs pairs. The cannot-link pairs are all the possible combinations consisting of one of the known ADR pairs and one of the known

Input :

- smoothing parameter $\mu > 0$ (e.g., 10^{-5})
- tolerance value tol (e.g., 10^{-5})
- step sized $\{\alpha_t \in (0,1) : t \in \mathbb{N}\}$

Output: $d \times d$ matrix $S_t^\mu \in S_+^d$
Initialization: $S_1^\mu \in S_+^d$ with $\mathbf{Tr}(S_1^\mu)= 1$
for $t = 1, 2, 3, ...$ **do**
  Z_t^μ=argmax$\{f_\mu(S_t^\mu) + \langle Z, \nabla f_\mu(S_t^\mu)\rangle : Z \in S_+^d$, **Tr**(Z)=1 $\}$, that is,
  $Z_t^\mu = \boldsymbol{\nu}\boldsymbol{\nu}^T$ where $\boldsymbol{\nu}$ is the maximal eigenvector of the matrix $\nabla f_\mu(S_t^\mu)$
  $S_{t+1}^\mu = (1 - \alpha_t)S_t^\mu + \alpha_t Z_t^\mu$
  if $|f_\mu(S_{t+1}^\mu) - f_\mu(S_t^\mu)| < tol$ then **break**
end

Algorithm 2: The distance metric learning algorithm from [211]

non-ADR pairs. It is then possible to apply the metric learning described by Algorithm 2 to efficiently learn the optimal metric space.

2.2.2.4 Semi-Supervised Learning Summary

In this section the frequently applied semi-supervised techniques have been summarised. The semi-supervised techniques can, under certain assumptions, improve classification/clustering by incorporating the unlabelled data when the number of labelled data are scarce [30]. Out of the semi-supervised classification techniques discussed (i.e, self-training, co-training and probability generating), the self-training algorithm is most applicable for classifying ADRs due to the probability generating algorithm requiring prior knowledge of the distributions [215], of which is unknown, and the difficulty with determining the non-correlated views required by the co-training algorithm.

Alternatively, the most suitable semi-supervised clustering technique is the

seed-constrained k-means [7] algorithm as this is efficient and takes advantage of the labelled data available. However, as discussed previously, clustering can be improved by applying metric learning [208]. As the ADR classification/clustering is required to be efficient, a suitable metric learning algorithm to apply to improve the clustering and ensure efficiency is the one presented in [211]. The choice of semi-supervised classification or semi-supervised clustering will need to be determined.

2.2.3 Pattern Recognition Summary

Statistical learning theory is a field of research that aims to learn or identify intrinsic patterns within data. These patterns can then be applied to make future predictions, and in the medical context, they can be used to aid decision making such as what drug to prescribe to a patient. When there are a sufficient number of labelled data, supervised learning can be applied whereby a general function is learned that accurately maps the input into the output. Numerous methods have been proposed that can produce a function that has a minimal training error but will also perform well on future data [39; 56; 81; 84; 101]. Ensemble methods have been presented that are able to combine multiple classifiers to reduce the variance and can improve the classification accuracy. Unfortunately, issues arise when using real life data. Such examples include the introduction of noise, difficulties generating labels or the presence of missing data. Ongoing research aims to develop methods that can produce an accurate function when there are issues present. In the case of insufficient labels, semi-supervised techniques have been proposed that make use of unlabelled data [215]. However, there is no guarantee

that semi-supervised algorithms will outperform their supervised counterpart [30]. Nonetheless, semi-supervised techniques have been successfully implemented on real life problems [108] and may be suitable for determining ADRs when there is a lack of known ADRs.

2.3 Literature Review Summary

The first section of the literature review focussed on the current techniques being applied by the pharmacovigilance community. The literature is full of techniques for signalling ADRs using LODs, but no method has been presented that was developed specifically for the THIN database and few studies have applied a range of methods on the THIN database. Therefore, there has been no extensive analysis of applying ADR signalling methods on the THIN database and a benchmark is unknown. The current research does highlight the inherent difficulties in accurately determining current benchmarks for ADR signalling techniques, and this will need to be addressed in order to find the THIN benchmark.

The majority of existing methods for signalling ADRs using LODs rely on measures of association strength or temporality and do not cover the seven other Bradford-Hill causality considerations [19]. Furthermore, they do not take into consideration attributes specific to the database being used, but database specific attributes may offer a unique insight into causality. As a consequence, the existing ADR signalling methods tend to be affected by confounding and this causes them to generate many false positives [156]. It may be possible to reduce the negative effects of confounding by generating attributes for each drug-medical event pair based on the remaining seven Bradford-Hill causality considerations

or by generating attributes specific to the database. The justification is that the Bradford-Hill causality considerations help distinguish between association and causation, something that is currently lacking within the existing methods. This may then result in a low false positive rate.

The existing ADR signalling techniques developed for LODs are unsupervised, as they do not learn from drug-medical event pairs that are known ADRs or non-ADRs. The reasoning being that it is difficult to obtain a large number of drug-medical event pairs with definitive ADR or non-ADR labels. However, if a sufficient number of ADR and non-ADR pairs were determined, then supervised or semi-supervised techniques could be applied, using suitable attributes, to identify new ADRs. The semi-supervised techniques may be advantageous when the number of labelled drug-medical event pairs are limited, for example in the case when a drug is rarely prescribed, then its ADRs may be generally unknown and the number of labelled data will be small.

A supervised or semi-supervised technique that uses attributes based on the Bradford-Hill causality considerations or specific to the THIN database may be able to reduce the negative effects of confounding by identifying and utilising patterns linked to ADRs or non-ADRs. The random forest ensemble algorithm is a suitable classifier to apply when there is a sufficient number of labelled data due to its ability to handle heterogeneous data and its resilience to noise. When the number of labelled data is low, a self-training algorithm or a semi-supervised clustering algorithm may yield improved results. If such an algorithm signalled ADRs with a low false positive rate, then a larger number of drug-medical event pairs likely to correspond to ADRs could be evaluated extensively by rigorous epidemiological studies, and this is likely to result in new ADRs being discovered

efficiently. In addition, the supervised/semi-supervised technique that considers more than just the association strength and temporality factors of the Bradford-Hill causality considerations (and may reduce the effects of confounding) is likely to outperform the existing ADR signalling methods on the THIN database and on the OMOP standard reference.

Chapter 3

Existing Methods Comparison

> 'One result from the DOI-HOI experiment was a number of reproducibly high false positive rates across methods and data sources.'
>
> Dubey et al. [48]

3.1 Introduction

So far in this thesis, the research hypotheses and aims have been defined and the current research within the field has been summarised. Numerous ADR signalling methods, specific for LODs, have been proposed, but few have been applied directly on the THIN database. As there has been no extensive application of existing methods applied on the THIN data, the general benchmark is unknown. As the aim of this research is to developed a suitable ADR signalling algorithm specifically for the THIN data, it is necessary to determine the current benchmark (i.e., the suitability of the existing methods on the THIN database).

In this chapter, the motivation for choosing two different types of comparison is given and followed by a description of the existing method implementations. The methodology used to determine the 'true' labels for each drug-medical event pair is proposed, as the signals generated by each method will be compared with the 'truth'. The various measures used to analyse each method's signalling ability are then presented and the comparisons are conducted. The chapter ends with a summary of the key results of both comparisons and the ADR signalling benchmark values for the THIN data are presented.

3.2 Motivation

ADRs are a consequence of multiple factors, for example, an ADR may only occur when the patient is a certain age and gender, eats a specific diet or has certain ongoing illnesses [94]. As a consequence, ADRs are difficult to identify and it is common for many ADRs to be unknown [147]. This means there is no extensive gold standard, as there is no complete list of definitive ADRs for any drug, and this makes it difficult to accurately benchmark ADR signalling algorithms. Motivated by the lack of gold standard, in [179], the authors developed a list of drug-medical event pairs known to represent ADRs or non-ADRs, but, although the list it expanding over time [70], it initially only considered four medical events. Further research has focused on producing a larger reference standard [33] containing drug-medical event pairs with definitive labels, but the number of drug-medical pairs is still often less than a hundred.

In previous studies, on non-THIN data, the authors have used the HOI-DOI reference standard containing 53 drug-medical event pairs with definitive labels,

and applied the existing methods to these pairs to determine how they compare and set an approximate benchmark [156]. In [156] the TPD and ROR_{05} signalled 16 drug-medical event pairs out a possible 53, with 6 and 4 known ADRs being signalled respectively. The benchmarks, over a range of electronic healthcare databases, for the TPD were an AUC of 0.73 and AP of 0.41 and for ROR_{05} an AUC of 0.68 and an AP of 0.2. The study also concluded that the existing methods obtain a similar performance and the existing methods have a high false positive rate, this was also evident in [162]. In the later study, the benchmark AUC obtained was 0.83.

A previous study determined the benchmark for the signalling ability of existing methods on the THIN database using the HOI-DOI reference standard [214]. The paper applied three existing methods, including the PRR and USCCS, to the THIN database mapped into the common data model and the raw THIN database. The results of the study, on the HOI-DOI reference standard, showed that the PRR and USCCS returned sensitivity values of 0.67 and 78 respectively and specificity values of 0.68 and 59 respectively on the mapped THIN database. Similar values were obtained by applying the PRR on the raw THIN data. Unfortunately, as the HOI-DOI reference standard restricts the analysis to a small subset of drug-medical event pairs, the impact of false positives is likely to be reduced (as there are less pairs to generate false positive on). This comparison may also add bias due to the choice of HOIs and DOIs included in the analysis. For example, the known ADRs included in the HOI-DOI reference standard have generally been signalled by numerous sources and may be easier to signal. Unfortunately, there has been no analysis of existing methods on the THIN database that includes a larger set of drug-medical event pairs, but this may be a more

realistic analysis.

To enable an extensive analysis of the signalling ability of the existing methods on the THIN database, additional comparisons with different perspectives and bias are required. The first perspective, referred to as the general comparison, generates signals using the existing methods for all the drug-medical event pairs satisfying the condition that the medical event occurs for at least one patient during the month after the drug. The true label for each drug-medical event pair is determined using current knowledge of ADRs, where only drug-medical event pairs currently known to be ADRs are considered true ADRs. Unfortunately, the known ADR status of each drug-medical event pair is not definitive, as some ADRs may be unknown, so this introduces error into the general comparison. The second perspective, referred to as the specific comparison, is similar to comparisons previously conducted [156], as it only analyses drug-medical event pairs that are either definitively non-ADRs or listed on drug packages as ADRs. However, the specific comparison considers a larger number of drug-medical event pairs than the HOI-DOI reference standard, so there may be less bias. The specific comparison is less affected by a lack of ADR knowledge than the general comparison, but may have errors due to drug package listed ADRs being potentially incorrect, due to the difficulty in determining causality.

In summary, as there is no gold standard, numerous comparisons need to be conducted to determine an extensive benchmark for the existing methods signalling ability on the THIN database. The HOI-DOI reference standard benchmark has been determined but this benchmark is limited due to potential bias caused by non-randomly selecting drug-medical event pairs. The general comparison will evaluate the methods without the selection bias, but will introduce

bias due to a lack of ADR knowledge. Finally, the specific comparison is a trade off between the previous comparisons and potentially contains bias from both non-random selection and a lack of ADR knowledge, but both types of bias are relatively reduced.

3.3 Existing Methods

To enable a fair comparison the TPD, MUTARA, HUNT and modified ROR methods, described in Chapter 2.1.4.2 were applied to investigate the one to thirty day period after the drug is prescribed (i.e., the month after). If each method used a different time period, the comparison would be biased.

3.3.1 TPD

In this study the TPD was implemented as described in [128], with IC value over the time period corresponding to the 30 days after the first prescription in 13 months ($u = [0, 30]$) contrasted with the IC value over the time period corresponding to the 27 to 21 months prior to prescription ($v = [-822, -639]$), but two different filters were investigated:

- The TPD is applied and medical events with an IC value the month prior to prescription or an IC value on the prescription day greater than the IC value during the month after the prescription are filtered (TPD 1).
- The TPD is applied and medical events with an IC value the month prior to prescription greater than the IC value during the month after the prescription are filtered (TPD 2).

The justification for choosing two filters is due to the possibility that ADRs can occur and be reported to doctors on the same day as the prescription, so filtering events with an IC value on the day of prescription greater than the IC value during the month after the prescription may prevent detection of some ADRs.

3.3.2 MUTARA & HUNT

Two different lengths for the reference period were investigated as the length of the reference period determines the per patient filter stringency and the optimal stringency is unknown for the THIN database. The reference period for MUTARA_{60} and HUNT_{60} is set to be the time period starting from two months prior to the prescription and ending the day before the prescription. The reference period for MUTARA_{180} and HUNT_{180} is set to be the time period starting from six months prior to the prescription and ending the day before the prescription. The reference periods are chosen to end the day before the prescribed as this gave better preliminary results. The other parameter values used are: $\text{T}_c = \text{T}_e = 30$, as this corresponds to the time period of a month after the drug prescription.

3.3.3 ROR

The 'Spontaneous reporting system' style transformation [216] is applied, where SRS style reports consisting of a patient, drug prescription and possible ADR are inferred from the LOD by discovering all the medical events that occur within 30 days of a drug prescription. Signals are only generated for medical events that have been reported with the drug of interest a minimum of 3 times.

3.4 Determining Labels

3.4.1 ADR Labels

The drug-medical event pair (α, β), consisting of a drug α and a medical event β, that correspond to an ADR were found using the online medical website NetDoctor [176] or using SIDER [98], a side effect resource containing side effects mined from drug packaging.

3.4.1.1 Online

The online medical website, NetDoctor, lists known ADRs for the majority of drugs available. The ADR strings for a general drug α were mined from the website. A string match was then applied to find the corresponding READ codes (e.g, SELECT READcode FROM Drugcodes WHERE description like '%ADR string%'), and each of the READ codes (β_is) that matched the NetDoctor listed ADRs were paired to the drug α and added to the set $\Psi^{\hat{A}}$,

$$\Psi^{\hat{A}} = \{(\alpha, \beta) | \beta \text{ is listed as an ADR to } \alpha \text{ on NetDoctor } \}$$

3.4.1.2 SIDER

The SIDER side effect resource contains information on drugs' ADRs and indications that were obtained by applying text mining to drug packaging. In total, the resource contains 996 drugs, 4192 ADRs and 99423 drug-medical event pairs corresponding to ADRs. The drug-medical event pairs, (α, β), corresponding to

ADRs were extracted from SIDER to generate the set Ψ^A,

$$\Psi^A = \{(\alpha, \beta)|\beta \text{ is listed as an ADR to } \alpha \text{ in SIDER}\}$$

3.4.2 Noise Labels

The noise labels were manually extracted by examining the THIN READ code tree. READ codes corresponding to irrelevant events such as 'Family history', 'Nationality', 'Job type', 'Chronic illnesses' (as this research is focusing on acute immediately occurring ADRs) or 'administrative events' were extracted and paired with all the drugs in the THIN database to generate the set Ψ^N,

$$\Psi^N = \{(\alpha, \beta)|\beta \text{ is irrelevant }, \alpha \in \text{ THIN}\}$$

3.5 Measures

For each drug, α, and medical event, β, the existing algorithms determine a measure of association between α and β. The TPD uses the $IC_{\Delta 05}(\alpha, \beta)$, MUTARA uses $unexlev(\alpha, \beta)$, HUNT uses the rank ratio, rank in descending order of $lev(\alpha, \beta)$ divided by rank in descending order of $unexlev(\alpha, \beta)$, and the modified SRS used the $ROR_{05}(\alpha, \beta)$ (reporting odds ratio lower 95% confidence interval). This measure of association is referred to as the rank score in the remained of this chapter.

Table 3.1: The signalling criteria of the existing methods

Method	Rank Score	Signal Criteria
TPD	$IC_{\Delta 05}$	$IC_{\Delta 05} > 0$
MUTARA	$unexlev$	$unexlev > 0$
HUNT	$\text{Rank}_{lev}/\text{Rank}_{unexlev}$	-
modified SRS	ROR_{05}	$ROR_{05} > 1$

Table 3.2: A worked example of comparing the existing methods signals and the known truth.

		Known Truth	
		ADR	Non-ADR
Signalled	Yes	True Positive $(TP) = 10$	False Positive $(FP) = 200$
	No	False Negative $(FN) = 12$	True Negative $(TN) = 500$

3.5.1 Natural Thresholds

The existing methods generate signals at their natural threshold, indicated in Table 3.1. The methods performances at their natural thresholds are generally uninformative as the natural threshold is an arbitrary cut off. However, in this thesis I will present the methods performances at their natural thresholds to enable comparison with existing work that has used these thresholds. To determine the method's ability to signal ADRs, the signals at the natural threshold are compared with the known truth. If a signalled drug-medical event pair is a true ADR then it's a True Positive, else it's a False Positive, conversely, if a non-signalled drug-medical event pair is a true ADR then it's a False Negative, else it's a True Negative, as summarised in Table 3.2. The measures of interest for the natural threshold can then be calculated as;

$$\text{Sensitivity} = TP/(TP + FN) \tag{3.1}$$

Table 3.3: An example of the medical event list associated to a specific drug and ordered by an existing method' rank score.

Medical Event	Rank Score	Known ADR	$y_{(i)}$	
Event 1	2.34	No	$y_{(1)} = 0$	
Event 5	2.12	Yes	$y_{(2)} = 1$	$P(2) = 1/2$
Event 4	1.75	Yes	$y_{(3)} = 1$	$P(3) = 2/3$
Event 2	1.74	No	$y_{(4)} = 0$	
Event 3	0.68	No	$y_{(5)} = 0$	

$$\text{Specificity} = TN/(TN + FP) \tag{3.2}$$

So, using the example in Table 3.2, the Sensitivity is $10/(10 + 12)$ and the specificity is $500/(500 + 200)$.

3.5.2 Ranking Ability

To determine the general ranking ability, each existing method is applied and returns a ranked list of the drug-medical event pairs being investigated in descending order of the rank score. Table 3.3 shows an example of the output of a method when considering the ranking of the medical events paired to the same drug. The function y_i, known as the truth, is 1 if the i^{th} ranked medical event is a known ADR and 0 otherwise. The precision of each method at cutoff K, denoted $P(K)$, is defined as the fraction of known ADRs that occur in the top K events of the list returned by each method for a specific drug, see Eq. (3.3).

$$P(K) = \frac{\sum_{i=1}^{K} y_{(i)}}{K} \tag{3.3}$$

Table 3.4: An example of the medical event list for all the drugs and ordered by one of the algorithms.

Drug	Medical Event	Rank Score	Known ADR	$y_{(i)}$
Drug 10	Event 7	2.34	No	$y_{(1)} = 0$
Drug 10	Event 5	2.12	Yes	$y_{(2)} = 1$
Drug 2	Event 56	1.75	Yes	$y_{(3)} = 1$
Drug 9	Event 7	1.74	No	$y_{(4)} = 0$
Drug 2	Event 16	0.68	No	$y_{(5)} = 0$

The average precision (AP) is a measure that can be used to determine how well a method generally ranks the medical events associated to a drug. This measure has previously been applied to compare methods implemented on the common data model [156]. The AP is calculated by finding the average $P(K)$ for each K corresponding to a known ADR,

$$\text{AP} = \frac{\sum_{K:y_{(K)}=1} P(K)}{\sum_i y_{(i)}} \tag{3.4}$$

Using Table 3.3 as an example, as there are two known ADRs returned ($\sum_i y_{(i)} = 2$) and the known ADRs in the table are ranked second and third we have $\{K : y_{(K)} = 1\} = \{2, 3\}$, so the AP score is,

$$\text{AP} = \frac{P(2) + P(3)}{2} = \frac{1/2 + 2/3}{2} = \frac{7}{12} \tag{3.5}$$

To give a general measure of the ranking ability of each algorithm over all the drugs investigated, the receiver operating characteristic (ROC) curves are computed. The ROC plots were generated by combining all the results for each method, as illustrated in Table 3.4. The ROC curves are formed by plotting the $sensitivity$ against $(1 - specificity)$. The Area Under the Curve (AUC) [28], was

approximated using the trapezoidal rule for a range of $specificity$ values ($AUC_{[a,b]}$ corresponds to the partial AUC [193] when only considering the $specificity$ within the interval $[a, b]$). To compare the AUCs of various methods DeLongs test at a 5% signifiance level is implemented [44].

3.6 General Comparison

3.6.1 Method

For the general comparison the method was as follows.

Step 1: Find the set of drug-medical event pairs such that the medical event is recorded within a $[1, 30]$ day time period after the drug for any patient. $G = \{(\alpha, \beta) | \beta$ occurs within the $[1, 30]$ day time interval centred around the day of the prescription of drug α for any patient $\}$.

Step 2: Determine the ground truth for each drug-medical event pair ($(\alpha, \beta) \in G$),

$$\text{Truth}(\alpha, \beta) = \begin{cases} \text{ADR}, & \text{if } (\alpha, \beta) \in \Psi^{\hat{A}} \\ \text{non-ADR}, & \text{otherwise} \end{cases} \tag{3.6}$$

Step 3: For each drug-medical event pair ($(\alpha, \beta) \in G$), calculate the method's rank score.

Step 4:

- Natural threshold- Determine signals using rank score and signal criteria. If (α, β) is signalled and $\text{Truth}(\alpha, \beta)$ is ADR then this is a TP, otherwise it is a FP. Conversely, if (α, β) is not signalled and $\text{Truth}(\alpha, \beta)$ is ADR then this is a FN, otherwise it is a TN.

Table 3.5: The specificity and sensitivity at the natural thresholds for the different algorithms (3dp).

Algorithm	Signals	Sens	Spec	Precision
$HUNT_{60}$	7785	0.179	0.903	0.0541
$HUNT_{180}$	7785	0.193	0.903	0.058
$MUTARA_{60}$	67624	0.933	0.109	0.032
$MUTARA_{180}$	65435	0.914	0.136	0.032
TPD 1	1893	0.090	0.953	0.057
TPD 2	3557	0.107	0.926	0.043
ROR_{05}	37729	0.312	0.726	0.031

- General Ranking - Plot the ROC curves and calculate the AUCs using the rank scores and Truth for each drug-medical event pair. The AP is also calculated on the list of medical events for each drug that are ordered in descending order of the assigned rank score, see Table 3.3.

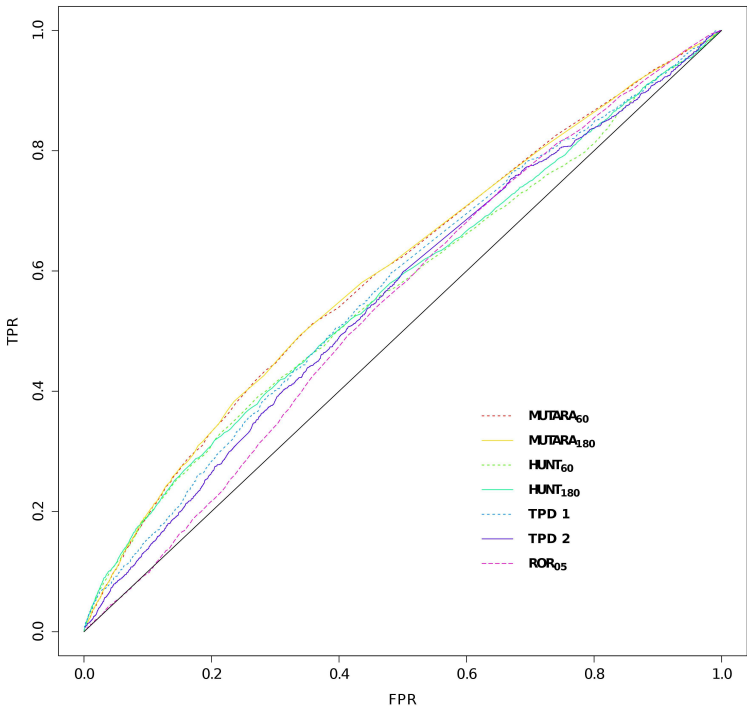
The existing methods were applied to 27 drugs for 6 drug families, for information about the drugs investigated, see Appendix B.

3.6.2 Results

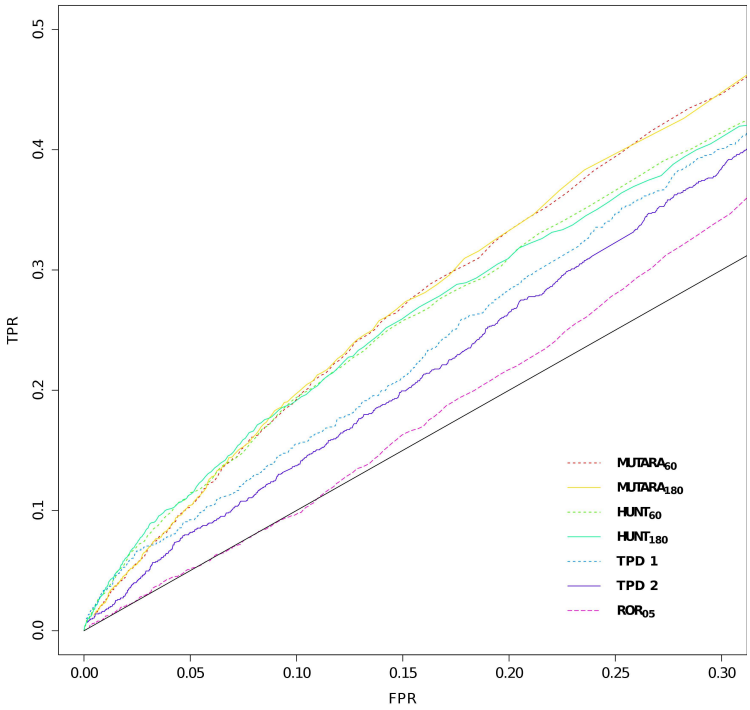
Table 3.5 shows the *specificity* and *sensitivity* for the different methods at their natural thresholds and the number of signals generated. As HUNT does not have a natural threshold, the top 10% of medical events were considered to be signalled.

The $AUC_{[0,1]}$ ranged between 0.546 (ROR_{05}) to 0.597 ($MUTARA_{180}$), the $AUC_{[0.7,1]}$ ranged between 0.048 (ROR_{05}) to 0.076 ($MUTARA_{60}$) and the $AUC_{[0.9,1]}$ ranged between 0.005 (ROR_{05}) to 0.011 ($HUNT_{180}$ and $HUNT_{60}$), as presented in Table 3.6. Figures 3.1a and 3.1b show the ROC plots for the different methods.

Figure 3.2a shows the AP scores for the different methods over the range



(a) Whole specificity range



(b) Section of specificity greater than 0.7

Figure 3.1: The ROC plots for the different methods. The black line is the line x=y.

Table 3.6: The AUC results for the different algorithms (3dp).

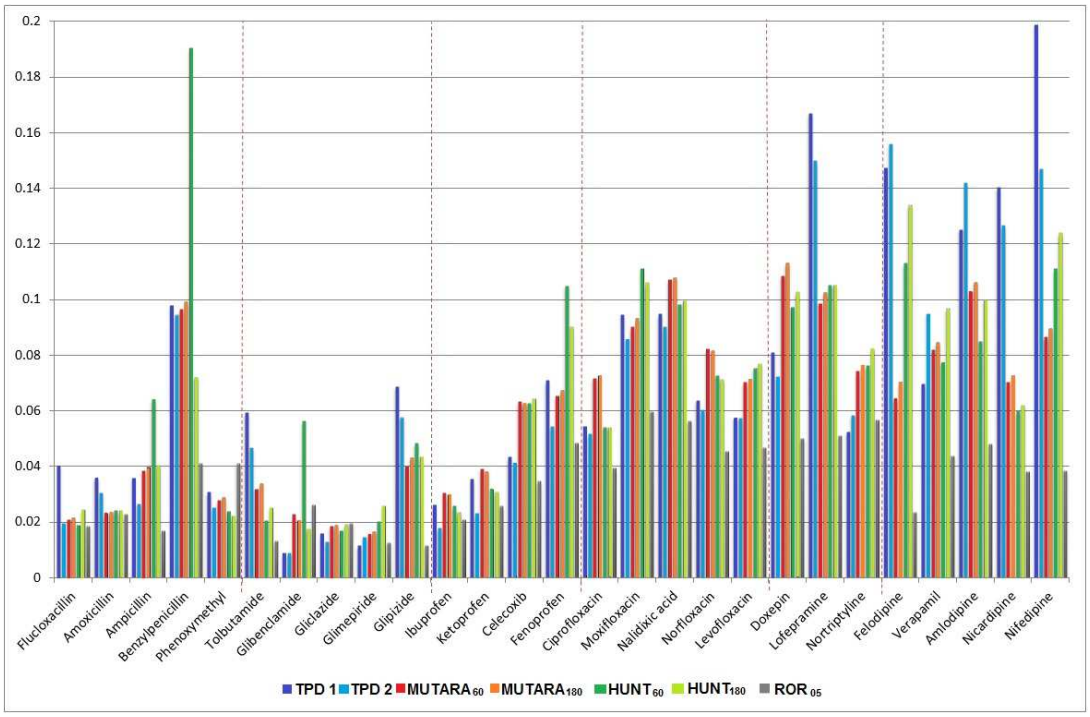
Algorithm	$AUC_{[0,1]}$	$AUC_{[0.7,1]}$	$AUC_{[0.9,1]}$
$HUNT_{60}$	0.566	0.072	0.011
$HUNT_{180}$	0.570	0.071	0.011
$MUTARA_{60}$	0.596	0.076	0.010
$MUTARA_{180}$	0.597	0.069	0.010
TPD 1	0.570	0.065	0.009
TPD 2	0.557	0.060	0.007
ROR_{05}	0.546	0.048	0.005

of drugs investigated. The family of drugs that the methods perform worse on overall were the sulphonylureas with AP scores ranging from $0.0088 - 0.0687$. The algorithms all performed well on the calcium channel blockers, with AP scores ranging from $0.0236 - 0.1988$, but the ROR_{05} performed worse for all the calcium channel blockers investigated. The methods also performed well for the tricyclic antidepressants with AP scores ranging between $0.0499 - 0.1670$. It can be seen in Figure 3.2a that generally the methods perform similarly between the same drugs of the same class, apart from the methods performing much better for benzylpenicillin sodium compared to the other penicillin drugs.

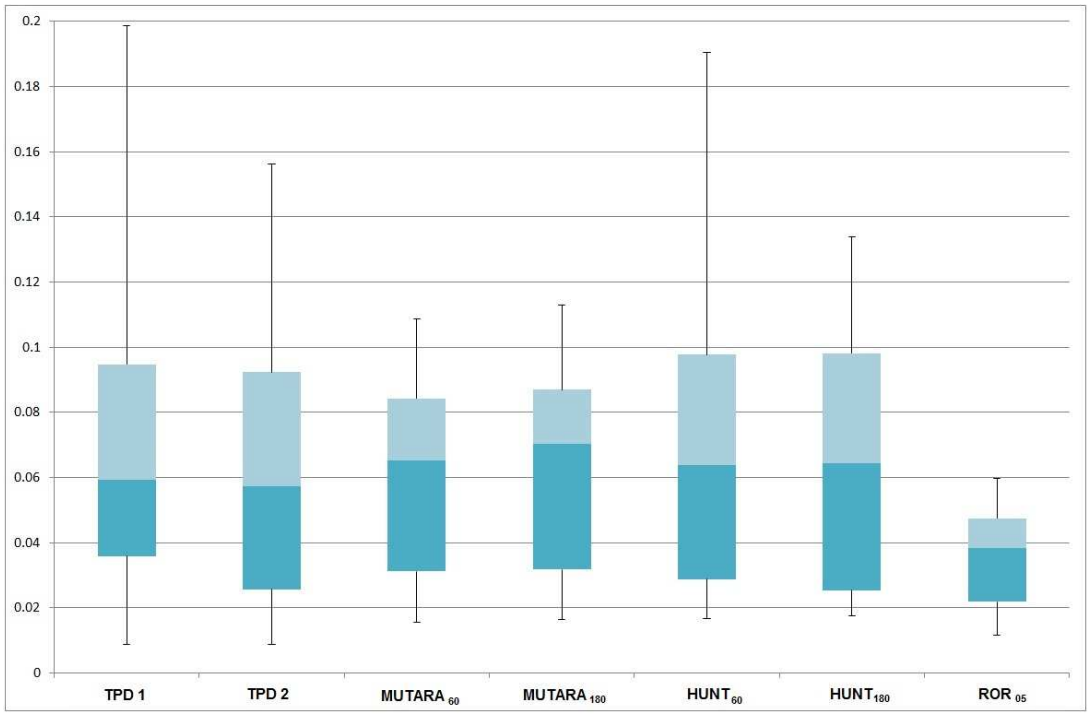
The box plots of the AP scores for the different methods seen in Figure 3.2b show overall the TPDs, MUTARAs and HUNTs perform equally and outperform the ROR_{05}. The MUTARA algorithm has the highest median AP score over all the drugs and is more consistent, whereas the performance of the TPD and HUNT varies more between the drugs.

3.6.3 Discussion

The results show that the methods' natural thresholds operate at different stringencies. The most stringent method was the TPD 1 that returned 1893 signals,



(a) Bar chart of the AP scores for each drugs.



(b) Box plot showing the median, quartiles and minimum/maximum AP scores.

Figure 3.2: AP results for each method applied for each drugs.

the lowest out of all the methods, with a high specificity of 0.953 and low sensitivity of 0.09, whereas the less stringent was the MUTARA_{60} that returned 67624 signals with a high sensitivity of 0.933 and a low specificity of 0.109. This was not unexpected as the TPD threshold used the lower confidence interval value rather than the actual IC_{Δ} value and the TPD applied a statistical shrinkage. The results also show that none of the methods was able to signal the known ADRs without being swamped by false positives signals.

The AUC results show that the methods perform similarly and no method had a higher partial AUC for all three restricted specificity intervals studied ($\text{AUC}_{[0,1]}$,$\text{AUC}_{[0.7,1]}$ and $\text{AUC}_{[0.9,1]}$). Overall no method consistently outperformed the others over all the drugs investigated in this study, however, either the TPD 1 or HUNT had the highest AP score for the majority of the drugs studied. The ROR_{05} generally performed the worse, but still had a higher AP score than the other methods for the drug phenoxymethylpenicillin.

The results obtained in this study were consistent with previous results as the $P(10)$ for MUTARA and HUNT averaged 0.065 and 0.122 respectively in this study and were 0.1 and $0.1 - 0.3$ respectively in previous work [91][90]. The $P(10)$ for the TPD method applied to Nifedipine in this study was 0.7, the same as on the UK IMS Disease Analyzer database [128]. However, there was deviation between the AP score of the ROR_{05} in this study $(0.01 - 0.06)$ and in the study by Zorych *et al.* [216] (0.1-0.15), this is probably due to this study using real data with redundant READ codes and Zorych *et al.* using simulated data. The general comparison also demonstrated that the existing methods generate a large number of false positive signals.

The main limitation of this comparative study was the assumption that if

a drug pair (α, β) is not in the set of known ADRs, Ψ^{A_1}, then it is a non-ADR. This is not true, as some pairs may be unknown ADRs, as there is no definitive complete list of ADRs for any drug. The consequence of this is that the true *sensitivity*, *specificity* and AP scores may be different to that the values obtained. However, the methods should be able to correctly rank the known ADRs and these are likely to be more common and obvious than the unknown ones. Therefore, if the method is unable to correctly rank the known ADRs above other events (and obtain a low AP in this study) then it is unlikely to identify the unknown ADRs, so the AP scores determined in this study still give insight into the methods abilities to detect ADRs. Another limitation was the READ code redundancy. The negative effect of noise may get amplified due to the redundancy causing there to be a larger number of noise READ codes. It may be the case that the methods would have higher AP scores if there was a way to group READ codes corresponding to the same medical event.

3.7 Specific Comparison

3.7.1 Method

Due to similar results being obtained in the general comparison by the MUTARA and HUNT methods implemented with different reference periods, only the MUTARA_{180} and HUNT_{180} were applied for the specific comparison. The method for the specific comparison is as follows.

Step 1: Find the definitive non-ADRs drug-medical event pairs corresponding to the drug of interest α or the drug-medical event pairs listed as ADRs on

α's drug packaging, $G = \{(\hat{\alpha}, \beta) \in \Psi^N \bigcup \Psi^A | \hat{\alpha} = \alpha\}$.

Step 2: Define the truth for each drug-medical event in G,

$$\text{Truth}(\alpha, \beta) = \begin{cases} \text{ADR}, & \text{if } (\alpha, \beta) \in \Psi^A \\ \text{non-ADR}, & \text{if } (\alpha, \beta) \in \Psi^N \end{cases} \tag{3.7}$$

Step 3: For each drug-medical event pair ($(\alpha, \beta) \in G$), calculate the method's rank score.

Step 4:

- Natural threshold- Determine signals using rank score and signal criteria. If (α, β) is signalled and Truth(α, β) is ADR then this is a TP, otherwise it is a FP. Conversely, if (α, β) is not signalled and Truth(α, β) is ADR then this is a FN, otherwise it is a TN.
- General Ranking - Plot the ROC curves and calculate the AUCs using the rank scores and Truth for each drug-medical event pair. The AP is also calculated on the list of medical events for each drug that are ordered in descending order of the rank score assigned by the existing method.

The existing methods were applied for five drugs, Nifedipine, Ciprofloxacin, Ibuprofen, Budesonide and Naproxen, see Appendix B.

3.7.2 Results and Discussion

For the specific comparison, MUTARA performed better at general ranking than the other methods, with greater AUC, $\text{AUC}_{[0.9,1]}$ and AP values, see Table 3.7. This result contradicts the general comparison results, that showed none of the

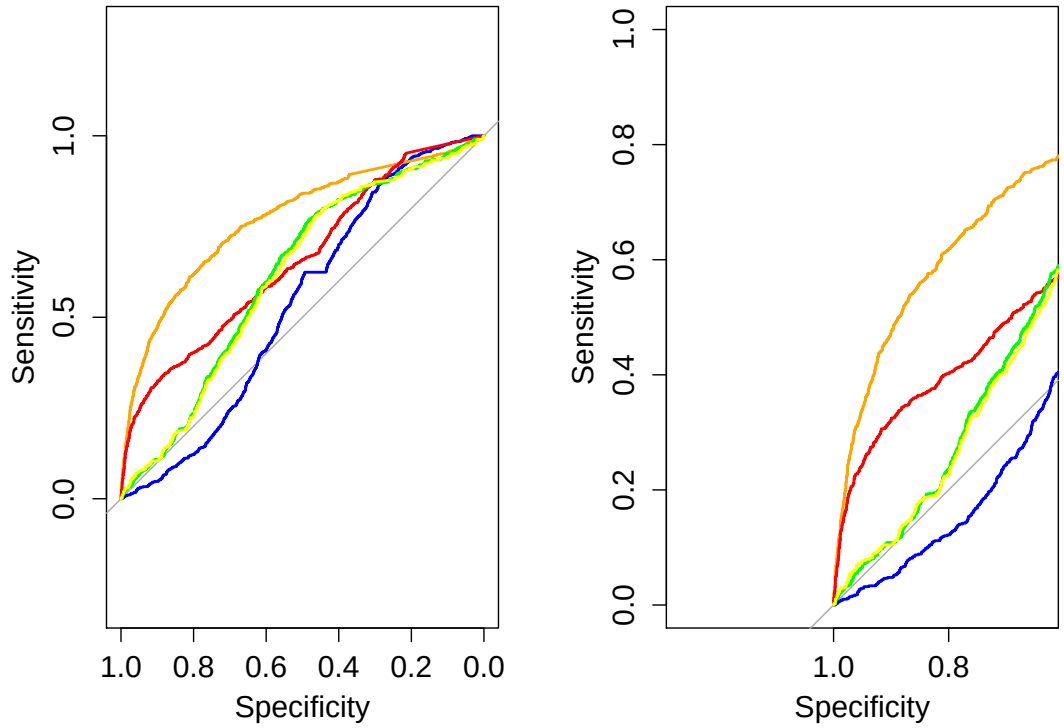


Figure 3.3: The ROC plots for the specific comparison. The figure on the left is the whole specificity range, the figure on the right is for the specificity within the interval $[0.9, 1]$. The orange, red, yellow, green and blue curves correspond to MUTARA_{180}, HUNT_{180}, TPD_1, TPD_2 and the ROR_{05} respectively.

Table 3.7: The ranking ability of the existing methods obtained in the specific comparison.

Method	AUC	$AUC_{[0.9,1]}$	AP
ROR_{05}	0.5374	0.003	0.072
$MUTARA_{180}$	0.770	0.032	0.315
$HUNT_{180}$	0.678	0.023	0.222
TPD_1	0.6149	0.007	0.095
TPD_2	0.6197	0.006	0.094

Table 3.8: The signals returned by the existing methods at their natural thresholds. The natural threshold used for HUNT was the rank ratio greater than 1.

Method	TP	FP	FN	TN	Sensitivity	Specificity	Precision
ROR_{05}	258	3197	429	4140	0.376	0.564	0.075
$MUTARA_{180}$	614	4648	73	2689	0.894	0.366	0.117
$HUNT_{180}$	466	4006	221	3331	0.678	0.454	0.104
TPD_1	42	302	645	7035	0.061	0.959	0.122
TPD_2	49	392	638	6945	0.071	0.947	0.111

existing methods outperforms any other when considering the overall ranking. However, the TPD was the method that returned the least number of false positives and obtained the greatest precision, 0.122. In agreement to previously obtained results, the specific comparison showed that the existing methods signal many false positives at their natural thresholds, see Table 3.8.

To identify why the TPD's ranking performance decreased relative to MUTARA for the specific comparison, the ranked list of drug-medical events pairs returned by the TPD was manually investigated. Interestingly, the manual investigation showed that the READ code redundancy was to blame, as the TPD did not assign a consistent $IC_{\Delta 05}$ for READ codes corresponding to the same medical event, and the READ codes matching the SIDER ADR strings tended to have lower $IC_{\Delta 05}$ values than other READ codes corresponding to known ADRs but

not exactly matching the SIDER ADR string (e.g. If a SIDER ADR string was 'vomiting' , then the THIN READ code with a description 'vomiting' would be labelled as an ADR, but 'O:E vomiting' or '[D] vomiting' would not be labelled). However, this does highlight that the TPD is not consistent and, although its performance may improve if different labels were used, it still struggles to assign a high $IC_{\Delta 05}$ to all READ codes corresponding to known ADR medical events . Previous studies have also identified inconsistency with the TPD [80].

The specific comparison appears to be a better way to compare the methods as the results are not limited by unknown ADRs. The potential bias introduced by only considering a subset of drug-medical event pairs has the advantage of highlighting methods that are not consistent. It can be argued that a perfect method would assign a similar rank score to READ codes corresponding to the same medical event, so methods unable to do this may be flawed.

3.8 Summary

In this chapter, four existing LOD ADR signalling methods were compared by applying them to the THIN database for a range of drugs. The comparisons measured how well they ranked the known ADRs or signalled known ADRs at their natural thresholds. As there is no golden standard, two different comparisons were applied. The first comparison compared the methods on a wide range of drug-medical event pairs but introduced bias by assuming there are no unknown ADRs, whereas the second comparison removed the bias of assuming there are no unknown ADRs but incorporated bias by only investigating a selection of drug-medical events pairs and by assuming drug packaging listed ADRs are correct.

The results highlight the issue of comparing existing methods without a golden standard. If a golden standard existed (i.e., for one drug all the ADRs were known), the methods could be applied to all the drug-medical event pairs for the specific drug and accurate measures could be obtained. However, when there is no gold standard, then bias is introduced and the results obtained may not be a true reflection of the methods abilities.

Nonetheless, considering the results of both comparisons and previous studies, the limitations of the existing methods were determined. The general comparison showed that no method was superior over all the drugs considered, however the specific comparison indicated that MUTATA is more consistently than the TPD. The main conclusion is that, for both comparisons, the existing methods failed to signal known ADRs without signalling a superfluous quantity of non-ADRs, resulting in a low precision benchmark of 0.122 for the specific comparison and 0.058 for the general comparison. The general ranking benchmarks for the general comparison are an AP of 0.2, an AUC of 0.597, an $AUC_{[0.9,1]}$ of 0.011 and an $AUC_{[0.7,1]}$ of 0.076. The benchmarks for the specific comparison are an AP of 0.315, an AUC of 0.770 and an $AUC_{[0.9,1]}$ of 0.032. Future methods should aim for higher scores.

Chapter 4

Incorporating Causation

> 'The application of Austin Bradford-Hill's criteria for evaluation causal associations in pharmacovigilance and pharmacoepidermiology is very useful.'
>
> Saad A.W. Shakir [163]

4.1 Introduction

So far in this thesis, the current research focus was summarised and the existing pharmacovigilance methodologies were presented. In the previous chapter the benchmark measures were determined by applying the current ADR signalling methods on the THIN database and it was concluded that they have a high false positive rate. In this chapter the processes implemented to generate and transform the data extracted from the THIN database are described. The main focus is the proposal of suitable attributes that offer insight into causality. The aim is to use these attributes as inputs into a learning algorithm that will be

trained to signal ADRs with a low false positive rate.

4.2 Motivation

The main limitation of ADR signal detecting using LODs is the abundance of confounding factors [68] [174]. The majority of medical events that occur after a drug are related to pre-existing illnesses, but these are still strongly associated to the drug. The existing methods can be considered unsupervised methods that aim to approximate the measure of causation between a drug-medical event pair. This is done by comparing the risk of the medical event after the drug compared with a substitute, such as the risk in a control population [91] or the risk when considering every other prescription [128]. Unfortunately, this only measures association as the choice of substitute introduces confounding [117], for example, as argued in [114], the choice of drug treatment may be influenced by the patient's medical history and the doctors preferences. To reduce the number of signals corresponding to medical events that are related to the drug cause, some authors have developed filters, such as ignoring medical events that occur more often before the drug than after [128; 161]. The consistently high false positive rate that occurs when the methods are applied to LODs suggests that these filters are still unable to removed all the effects of confounding and this hinders the effectiveness of the existing methods. The signals they generate require further analysis [148] and rare ADRs may not be signalled [143].

To develop an improved ADR signalling algorithm, it is important to identify a way to distinguish between association and causation in observational data. Such an algorithm would have a reduced false positive rate as it would be resilient

against confounding effects. Causality is often determined using a randomised controlled experiment [154], where treatments and controls are randomly assigned to control confounding [153]. This cannot be implemented using observational data, as there is no control over who is assigned a treatment. In [67], the authors highlight the issues associated with using observational data for causal inference. A common technique to identify causality using observational data is to apply a supervised algorithm with additional knowledge that incorporates confounding factors, such as fitting a regression model that incorporates parameters based on confounding [58]. In [132] and [42] the authors manually investigated causality between a single drug and a single medical event by investigating suitable measures of the Bradford-Hill causality considerations that can be derived from pharmacovigilance data. The Bradford-Hill causality considerations are often used when determining causation [164] and researchers have discussed the importance of applying these considerations within pharmacovigilance [52]. In this thesis, the idea is expanded by removing the requirement of a manual inspection. Instead, an algorithm is implemented that learns to determine causality between each drug-medical event pair based on Bradford-Hill causality derived attributes, as this will increase efficiency and enable a wider search. The novel idea is to train a supervised algorithm using attributes based on latent variables (not directly observed), derived from the Bradford-Hill causality considerations, rather than observable variables. In this chapter, the attributes that add insight on causality, derived from a selection of the Bradford-Hill causality considerations, are proposed and explored.

4.2.1 Data Cleansing

The THIN database contains validation fields indicating the integrity of each record. Only records that are valid are extracted from the database and records corresponding to patients with a missing date of birth or gender are deleted. Any records containing an invalid age such as a negative number or greater than one hundred and twenty years are also removed.

Patients whom are newly registered present a problem in the THIN database as they often come to the practice with a magnitude of historical and existing conditions that get recorded during their first few visits even when the conditions initially occurred years previously. This is often referred to as 'registration event dropping'. Studies have shown that the probability of 'registration event dropping' is significantly reduced after a patient has been registered at the same practice for a year [104]. To prevent this biasing the results the medical records that were recorded within the first year of each patient being registered are deleted from the THIN database. The last month of prescription records for patients are also ignored to prevent under-reporting, as implemented in [143].

4.2.2 Data Extraction

4.2.2.1 Formulation

Denoting the cleansed THIN data by $\Omega = \{\Omega^P, \Omega^E\}$, where Ω^P is the set of valid drug prescription reports and Ω^E is the set of valid medical event reports contained in the THIN database. Throughout this thesis, the i^{th} element of the vector $\mathbf{x}$ is represented by x_i. Each prescription record, $\boldsymbol{\omega} \in \Omega^P \subset \mathbb{R}^8$, is a vector containing the details about the prescription, where $\boldsymbol{\omega} \in \mathbb{R}^8$ and,

ω_1 : is the corresponding record's patient ID (who had the prescription).

ω_2 : is the corresponding record's prescribed drug.

ω_3 : is the corresponding record's gender (1 if male, 0 if female).

ω_4 : is the corresponding record's date of prescription (when the prescription was issued).

ω_5 : is the corresponding record's patient's age (time in days between the patients date of birth and when the prescription was issued).

ω_6 : is the corresponding record's dosage (dosage of the drug issued).

ω_7 : is the corresponding record's British National Formulary (BNF) code (a code specifying the general family of the drug).

ω_8 : is the corresponding record's noise value (the total number of drugs prescribed within the $[-30, 30]$ day time interval centred around the prescription date).

Each medical event record in the complete THIN database, $\boldsymbol{\psi} \in \Omega^E \subset \mathbb{R}^5$, is a vector containing the medical event report details, where $\boldsymbol{\psi} \in \mathbb{R}^5$ and,

ψ_1 : is the corresponding record's patient ID (who had the medical event).

ψ_2 : is the corresponding record's READ code (corresponding to medical event).

ψ_3 : is the corresponding record's date of recording (when the medical event was reported).

ψ_4 : is a binary value representing if a READ code with the same Level 3 READ code parent as the record has been recorded for the patient before. If it is the first time the value is 1, otherwise it is 0.

ψ_5 : is a binary value representing if a READ code with the same Level 4 READ code parent as the record has been recorded for the patient before. If it is the first time the value is 1, otherwise it is 0.

As it can be seen, each medical event is recorded via a READ code. The READ codes have a tree structure with five levels of specificity, as described in Appendix A. Therefore, the term drug-medical event pair is analogous to the term drug-READ code pair. For clarity, the drug-medical event pair corresponding to drug α and READ code β is denoted by (α, β). Unfortunately, the READ codes are redundant, and multiple READ codes can correspond to the same medical event but with slight variance in the description. For example, the READ code '91a..' may represent 'had a chat with patent' and the READ code '91b..' may represent 'discussion with patient at his request', both these READ codes correspond to the medical event of talking to the patient, but differ slightly.

4.2.2.2 Extraction

For a given drug, α, it is possible to extract prescription records of interest in three different ways. The first method extracts all the records containing the drug α (where ω_2 was previously denoted as the drug prescribed in therapy record $\boldsymbol{\omega}$),

$$\Omega^{P_\alpha} := \{\boldsymbol{\omega} \in \Omega^P | \omega_2 = \alpha\} \tag{4.1}$$

In the latter part of this chapter the prescription records in the set Ω^{P_α} are used to find a rough measure of association, by investigating the medical events that occur shortly before the drug compared with the medical events that occur shortly after a drug. As there is no restriction on how far apart the prescription records in the set Ω^{P_α} are for the same patient, some prescriptions, for the same patient, may be recorded in short succession. This may cause bias when investigating the medical events that occur shortly before one prescription, as they may be caused

by the previous prescription.

To reduce this bias a second method to extract reports containing drug α is proposed. For two therapy records, $\boldsymbol{\omega}, \overset{*}{\boldsymbol{\omega}} \in \Omega^{P_\alpha}$, the records correspond to the same patient when both patient IDs are the same, $\omega_1 = \overset{*}{\omega}_1$, and correspond to the same drug when $\omega_2 = \overset{*}{\omega}_2$. The second method only extracts reports if the drug was not previously prescribed, to the same patient, within the previous 13 months,

$$\hat{\Omega}^{P_\alpha} := \{\boldsymbol{\omega} \in \Omega^{P_\alpha} | \ \{\overset{*}{\boldsymbol{\omega}} \in \Omega^{P_\alpha} \setminus \boldsymbol{\omega} | \omega_1 = \overset{*}{\omega}_1, \omega_2 = \overset{*}{\omega}_2, t_m(\overset{*}{\omega}_4, \omega_4) \in [0, 13]\} = \emptyset\}$$

Where the function $t_m(a, b) : \text{Date} \times \text{Date} \to \mathbb{Z}$ denotes the time in months from date a to date b. As different drugs from the same family are often prescribed, due to the first drug not being effective or the patient reacting badly, there can still be bias when using $\hat{\Omega}^{P_\alpha}$. To further reduce the bias, only prescription records where there has been no previous prescription of a similar drug within the previous 13 months are considered,

$$\overline{\Omega}^{P_\alpha} := \{\boldsymbol{\omega} \in \Omega^{P_\alpha} | \ \{\overset{*}{\boldsymbol{\omega}} \in \Omega^{P} \setminus \boldsymbol{\omega} | \omega_1 = \overset{*}{\omega}_1, \omega_7 = \overset{*}{\omega}_7, t_m(\overset{*}{\omega}_4, \omega_4) \in [0, 13]\} = \emptyset\}$$

Figure 4.1 is a graphic representation of the different filtering that is implemented to extract the data, where a drug is considered filtered if it is surrounded by a red square. Each line represents the sequence of drug records ordered by time, where drug 1 and 2 are from the same drug family (have the same BNF code). The top line represents no filtering, so all drug records are included in the analysis, the middle line represents filtering a drug if the same drug was recorded

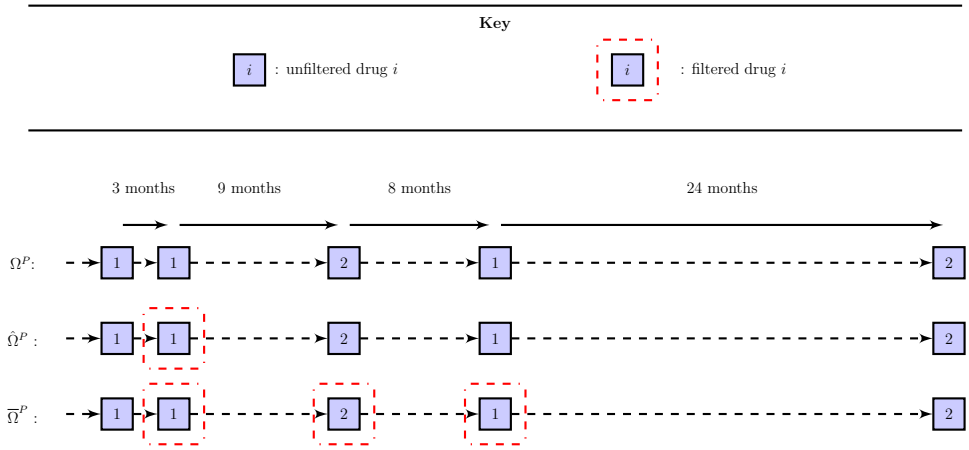


Figure 4.1: Illustration of filtering done during data extraction.

within 13 months previously, and the bottom line represents filtering a drug if a drug in the same family was recorded within the 13 months previously. Using the prescription record subsets for the drug α, it is possible to determine the 'risk' drug-READ code pairs that correspond to potential acutely occurring ADRs by finding all the READ codes that occur within 30 days of any prescription of α, see Equation 4.2. Figure 4.2 illustrates how the 'risk' drug-READ code pairs corresponding to drug 1 are determined, where the medical events represented by circles are paired with drug 1 if they occur between the square representing drug 1 and the red line. A short time period is used as the focus of this research is on discovering ADRs that occur immediately after ingesting a drug. It was determined that investigating the 30 days after a drug is prescribed was the best trade off between having a sufficiently large period of time after the prescription to allow patients time to report the medical event while not having the time period too large that many erroneous medical events will be reported.

$$RME^{\alpha} = \{\psi_2 | \boldsymbol{\psi} \in \Omega^E, \exists \boldsymbol{\omega} \in \Omega^{P_\alpha} \text{ where } \omega_1 = \psi_1, t_d(\omega_4, \psi_3) \in [1, 30]\} \quad (4.2)$$

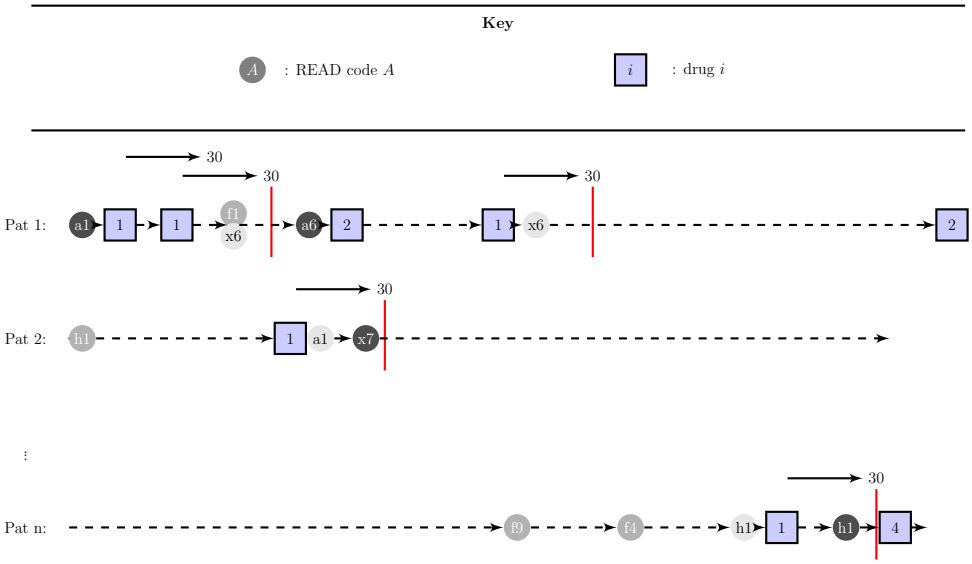


Figure 4.2: Illustration of determining risk drug-medical event pairs.

where the function $t_d(a, b) : \text{Date} \times \text{Date} \to \mathbb{Z}$ represents the time in days from date a to date b. One suitable approach to determine ADRs to drug α using LODs is to combine medical event reports with prescription reports containing α when the medical event report occurs within a set time frame around the prescription report. Let $\Omega^{[u,v],P_\alpha}$ be a relationship between the prescription records of drug α (Ω^{P_α}) and the medical event records (Ω^E) defined by the records having the same patient ID ($\omega_1 = \psi_1$) and the medical event report occurring within the set time period around the date of the prescription report ($t_d(\omega_4, \psi_3) \in [u, v]$),

$$\Omega^{[u,v],P_\alpha} = \{(\boldsymbol{\omega}, \boldsymbol{\psi}) \in \Omega^{P_\alpha} \times \Omega^E | \omega_1 = \psi_1, t_d(\omega_4, \psi_3) \in [u, v]\} \tag{4.3}$$

As illustrated in Equation (4.3), each element in the set $\Omega^{[u,v],P_\alpha}$ contains the combined prescription reports containing α and medical event reports of interest, where the medical event report occurred with the $[u, v]$ day interval around the

prescription record. For each combined record, $\boldsymbol{\kappa} = (\boldsymbol{\omega}, \boldsymbol{\psi}) \in \Omega^{[u,v],P_\alpha}$, the first eight elements correspond to the prescription record ($\kappa_i = \omega_i, i \in [1,8]$) and the last five elements correspond to the medical event record ($\kappa_i = \psi_{i-8}, i \in [9,13]$). Similarly, the set of combined reports of interest can be calculated when only considering the first time prescriptions in 13 months of drug α or the prescriptions of drug α that have no similar drug prescribed within the previous 13 months by substituting the set Ω^{P_α} with the set $\hat{\Omega}^{P_\alpha}$ or $\overline{\Omega}^{P_\alpha}$ respectively,

$$\hat{\Omega}^{[u,v],P_\alpha} = \{(\boldsymbol{\omega}, \boldsymbol{\psi}) \in \hat{\Omega}^{P_\alpha} \times \Omega^E | \omega_1 = \psi_1, t_d(\omega_4, \psi_3) \in [u,v]\} \tag{4.4}$$

$$\overline{\Omega}^{[u,v],P_\alpha} = \{(\boldsymbol{\omega}, \boldsymbol{\psi}) \in \overline{\Omega}^{P_\alpha} \times \Omega^E | \omega_1 = \psi_1, t_d(\omega_4, \psi_3) \in [u,v]\} \tag{4.5}$$

The aim of this thesis is to develop a classifier such that, for each 'risk' drug-READ code pair containing drug α, (α,$\beta \in RME^\alpha$), the pair is classified as either an ADR or non-ADR. To develop such a classifier requires generating suitable attributes for each pair and learning from pairs that are known ADRs and non-ADRs. In the next section the proposed attributes based on the Bradford-Hill causality considerations and THIN structures are derived.

4.2.3 Data Derivation

After cleansing and extracting the data of interest, the data can now be transformed into suitable attributes. The set of combined reports containing READ code β that occur within the $[u,v]$ time interval centred around the prescription

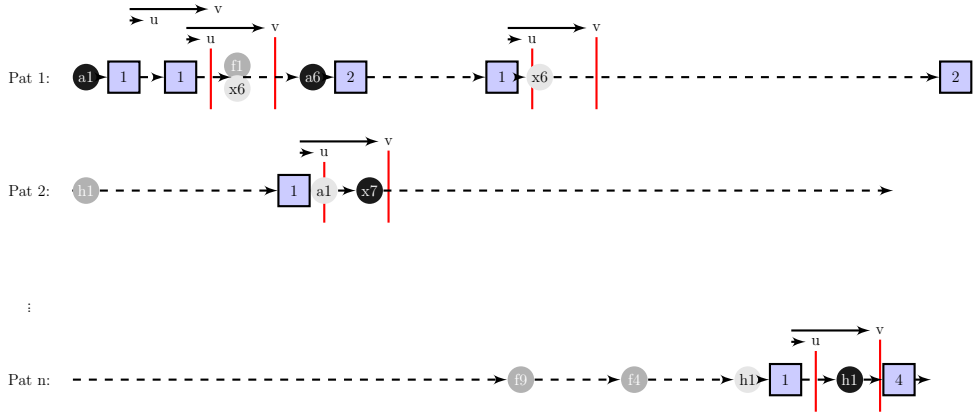


Figure 4.3: Illustration of combining drug records and medical event records.

of drug α is,

$$\Omega^{[u,v],P_\alpha E_\beta} = \{\boldsymbol{\kappa} \in \Omega^{[u,v],P_\alpha} | \kappa_{10} = \beta\} \tag{4.6}$$

The equivalent sets when only considering the prescriptions of drug α for the first time in 13 months or prescriptions of drug α when a similar family drug has not been prescribed within the last 13 months are,

$$\hat{\Omega}^{[u,v],P_\alpha E_\beta} = \{\boldsymbol{\kappa} \in \hat{\Omega}^{[u,v],P_\alpha} | \kappa_{10} = \beta\} \tag{4.7}$$

and

$$\overline{\Omega}^{[u,v],P_\alpha E_\beta} = \{\boldsymbol{\kappa} \in \overline{\Omega}^{[u,v],P_\alpha} | \kappa_{10} = \beta\} \tag{4.8}$$

respectively.

This is graphically illustrated in Figure 4.3, where the medical event reports are represented by circles and the drugs are represented by squares. For each prescription of drug 1, the time interval $[u, v]$ centred around the prescription is investigated, and any medical event report occurring within the interval is combined with the report to produce a new combined report.

Table 4.1: Contingency table calculations for the epidemiological association measures.

	READ code β	not READ code β
Drug α	$\lvert\Omega^{[1,30],P_\alpha E_\beta}\rvert$	$\lvert\Omega^{[1,30],P_\alpha}\rvert - \lvert\Omega^{[1,30],P_\alpha E_\beta}\rvert$
$\bigcup_{\gamma\neq\alpha}$Drug$_\gamma$	$\sum_{\gamma\neq\alpha} \lvert\Omega^{[1,30],P_\gamma E_\beta}\rvert$	$\sum_{\gamma\neq\alpha} \lvert\Omega^{[1,30],P_\gamma}\rvert - \sum_{\gamma\neq\alpha} \lvert\Omega^{[1,30],P_\gamma E_\beta}\rvert$

4.2.3.1 Association Strength

> 'Prospective inquiries into smoking have shown that the death rate for cancer of the lung in cigarette smokers is nine to ten times the rate in non-smokers' - **Bradford-Hill** [19].

The association strength criterion concentrates on how associated the READ code and drug are [19]. Previously implemented measures of association derived from LODs are the IC$_\Delta$ used by the TPD algorithm [128], see Chapter 2.1.4.2, or the IC$_{\Delta 05}$. These values measure the association between the READ code and drug during the 'hazard' period that occurs after the prescription relative to some 'non-hazard' time period.

$$x_1^{\alpha\beta} = IC_\Delta \tag{4.9}$$

$$x_2^{\alpha\beta} = IC_{\Delta 05} \tag{4.10}$$

Alternative measures are the risk ratio (RR), odds ratio (OR) and risk different (RD) that are frequently used in epidemiological studies to measure the association between an exposure and disease [194]. These measures contrast the rate of disease in a population that is exposed to some risk with the rate of the disease in a population not exposed. In [151] the author states that, if there are other (non-drug) sufficient causes of a medical event, then the frequency of these

has a greater impact on the medical event and drug's risk ratio measure than the risk difference measure. So the measures can generate varying strengths of association for the same drug-READ code pair.

The RD investigates the additive difference between the risk of having the READ code after the drug of interest is prescribed compared to the risk of having the READ code after any other drug prescription [181]. Using Table 4.1 the RD can be calculated by,

$$\begin{aligned} x_3^{\alpha\beta} &= (|\Omega^{[1,30],P_\alpha E_\beta}|/|\Omega^{[1,30],P_\alpha}|) - (\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma}|) \\ x_4^{\alpha\beta} &= (|\hat{\Omega}^{[1,30],P_\alpha E_\beta}|/|\hat{\Omega}^{[1,30],P_\alpha}|) - (\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma}|) \\ x_5^{\alpha\beta} &= (|\overline{\Omega}^{[1,30],P_\alpha E_\beta}|/|\overline{\Omega}^{[1,30],P_\alpha}|) - (\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma}|) \end{aligned} \tag{4.11}$$

The RR estimates the risk of having the READ code in the month after the drug of interest is prescribed divided by the risk of having the READ code in the month after any other drug, the measure has been incorporated to signal ADRs in SRS databases [9]:

$$\begin{aligned} x_6^{\alpha\beta} &= (|\Omega^{[1,30],P_\alpha E_\beta}|/|\Omega^{[1,30],P_\alpha}|)/(\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma}|) \\ x_7^{\alpha\beta} &= (|\hat{\Omega}^{[1,30],P_\alpha E_\beta}|/|\hat{\Omega}^{[1,30],P_\alpha}|)/(\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma}|) \\ x_8^{\alpha\beta} &= (|\overline{\Omega}^{[1,30],P_\alpha E_\beta}|/|\overline{\Omega}^{[1,30],P_\alpha}|)/(\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma E_\beta}|/\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma}|) \end{aligned} \tag{4.12}$$

The OR calculates the ratio between the odds that a READ code occurs in the

drug of interest group and the odds that a READ code occurs in any other drug group. There has been much debate into the usefulness of this measure, in [165] the authors state that the OR should not be used in place of the RR although in [194] they argue that the OR and RR are different measures and as long as the OR is not considered on the same scale as the RR then it is worthwhile:

$$
\begin{aligned}
x_9^{\alpha\beta} &= \left(\frac{|\Omega^{[1,30],P_\alpha E_\beta}|}{[|\Omega^{[1,30],P_\alpha}| - |\Omega^{[1,30],P_\alpha E_\beta}|]}\right)/\left(\frac{\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma E_\beta}|}{[\sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma}| - \sum_{\gamma\neq\alpha}|\Omega^{[1,30],P_\gamma E_\beta}|]}\right) \\
x_{10}^{\alpha\beta} &= \left(\frac{|\hat{\Omega}^{[1,30],P_\alpha E_\beta}|}{[|\hat{\Omega}^{[1,30],P_\alpha}| - |\hat{\Omega}^{[1,30],P_\alpha E_\beta}|]}\right)/\left(\frac{\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma E_\beta}|}{[\sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma}| - \sum_{\gamma\neq\alpha}|\hat{\Omega}^{[1,30],P_\gamma E_\beta}|]}\right) \\
x_{11}^{\alpha\beta} &= \left(\frac{|\overline{\Omega}^{[1,30],P_\alpha E_\beta}|}{[|\overline{\Omega}^{[1,30],P_\alpha}| - |\overline{\Omega}^{[1,30],P_\alpha E_\beta}|]}\right)/\left(\frac{\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma E_\beta}|}{[\sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma}| - \sum_{\gamma\neq\alpha}|\overline{\Omega}^{[1,30],P_\gamma E_\beta}|]}\right)
\end{aligned}
\tag{4.13}
$$

4.2.3.2 Temporality

> 'Does a particular diet lead to disease or do the early stages of the disease lead to those peculiar dietetic habits?' - **Bradford-Hill** [19].

The temporality criteria concerns itself with the direction of the relationship between the READ code and drug. This is an important factor, and has been highlighted in other criteria for causation [86]. It measures if the READ code occurs after the drug, or the other way round. If the READ code and drug are associated but the READ code frequently occurs before the drug, then this may suggest the medical event corresponding to the READ code causes the drug and not the other way round.

The first values of interest are the after and before ratios (AB ratios). The

AB ratios calculate how many prescriptions of α have the READ code β recorded between 1 and 30 days after the prescription divided by how many have the READ code β recorded between 1 and 30 days before the prescription, this is a basic implementation of the self controlled cross-over method.

$$\begin{aligned} x_{12}^{\alpha\beta} &= |\Omega^{[1,30],P_\alpha E_\beta}|/|\Omega^{[-30,-1],P_\alpha E_\beta}| \\ x_{13}^{\alpha\beta} &= |\hat{\Omega}^{[1,30],P_\alpha E_\beta}|/|\hat{\Omega}^{[-30,-1],P_\alpha E_\beta}| \\ x_{14}^{\alpha\beta} &= |\overline{\Omega}^{[1,30],P_\alpha E_\beta}|/|\overline{\Omega}^{[-30,-1],P_\alpha E_\beta}| \end{aligned} \tag{4.14}$$

Other suitable attributes for the temporality criterion are the filters that have been implemented by existing LOD ADR signalling algorithms, where $x_{15}^{\alpha\beta}$ and $x_{16}^{\alpha\beta}$ correspond, respectively, to the TPD Filter 1 and the TDP Filter 2 , the modified versions of the TPD filter applied initially in [128] and adapted in [143]. The final attribute, $x_{17}^{\alpha\beta}$, is the output of the LEOPARD algorithm described in [161].

4.2.3.3 Specificity

> 'If the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favour of causation' - **Bradford-Hill** [19].

The third Bradford-Hill consideration is how specific an association is. In general, specificity is interpreted as the drug only causes a single, specific, ADR [67]. Consequently, many researchers, including [151] and [67], argue this is not very informative, as many drugs cause multiple ADRs . Other researchers have

suggested modifying the interpretation of the specificity criteria [197]. Weiss argues that an association is specific when both the outcome and the exposure are specific or when only specific people that are exposed have the outcome.

This prompts the novel specificity attributes proposed in this thesis. The first considers how specific the READ code is, justification for this is that general outcomes are likely to occur by chance as they probably have a high background rate, but if a specific READ code occurs frequently after the drug of interest is prescribed then this may give reason to suspect it as an ADR. The first specificity attribute uses the READ code hierarchal level,

$$x_{18}^{\alpha\beta} = i, \text{ where } \beta \text{ corresponds to a level } i \text{ READ code} \tag{4.15}$$

If the drug and READ code association is only found in a certain subpopulation then this may also be suggestive of an ADRs. Therefore attributes are developed based on the age and gender of the patients experiencing the READ code after the drug compared to all the patients who are prescribed the drug. A suitable method to measure if a specific age group experience β after α is to compare the average age of the patients who experience β within 1 to 30 days after α with the

average age of the patients prescribed α.

$$
\begin{aligned}
x_{19}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\Omega^{[1,30],P_\alpha E_\beta}} \kappa_5/|\Omega^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\Omega^{[1,30],P_\alpha}} \kappa_5/|\Omega^{[1,30],P_\alpha}|\Big) \\
x_{20}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\hat{\Omega}^{[1,30],P_\alpha E_\beta}} \kappa_5/|\hat{\Omega}^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\hat{\Omega}^{[1,30],P_\alpha}} \kappa_5/|\hat{\Omega}^{[1,30],P_\alpha}|\Big) \\
x_{21}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\overline{\Omega}^{[1,30],P_\alpha E_\beta}} \kappa_5/|\overline{\Omega}^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\overline{\Omega}^{[1,30],P_\alpha}} \kappa_5/|\overline{\Omega}^{[1,30],P_\alpha}|\Big)
\end{aligned}
\tag{4.16}
$$

Justified by a similar argument, it is also useful to calculate a measure to compare the ratio of patients that experience β within 1 and 30 days of α that are male relative to the ratio of patients who are prescribed α and are male.

$$
\begin{aligned}
x_{22}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\Omega^{[1,30],P_\alpha E_\beta}} \kappa_3/|\Omega^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\Omega^{[1,30],P_\alpha}} \kappa_3/|\Omega^{[1,30],P_\alpha}|\Big) \\
x_{23}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\hat{\Omega}^{[1,30],P_\alpha E_\beta}} \kappa_3/|\hat{\Omega}^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\hat{\Omega}^{[1,30],P_\alpha}} \kappa_3/|\hat{\Omega}^{[1,30],P_\alpha}|\Big) \\
x_{24}^{\alpha\beta} &= \Big(\sum_{\boldsymbol{\kappa}\in\overline{\Omega}^{[1,30],P_\alpha E_\beta}} \kappa_3/|\overline{\Omega}^{[1,30],P_\alpha E_\beta}|\Big)/\Big(\sum_{\boldsymbol{\kappa}\in\overline{\Omega}^{[1,30],P_\alpha}} \kappa_3/|\overline{\Omega}^{[1,30],P_\alpha}|\Big)
\end{aligned}
\tag{4.17}
$$

4.2.3.4 Biological Gradient

> 'The fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simple evidence that cigarette smokers have a higher death rate than non-smokers.' - **Bradford-Hill** [19].

The biological gradient criterion in the context of ADR detection considers the

dosage of the drug. Often, but not always the case, an ADR is more likely to occur when the drug is ingested at a high dosage compared to a low dosage [36]. In [192] the authors used the Pearson's correlation and logistical regression to measure the biological gradient. However, in [142], it was shown that the Pearson's correlation was difficult to calculate using data from the THIN database. Therefore different measures are required.

The proposed novel biological attribute, calculated below, contrasts the average drug dosage for the patients that experience β within 1 to 30 days of α with the average drug dosage for all the patients prescribed α,

$$x_{25}^{\alpha\beta} = (\sum_{\boldsymbol{\kappa} \in \overline{\Omega}^{[1,30],P_\alpha E_\beta}} \kappa_6 / |\overline{\Omega}^{[1,30],P_\alpha E_\beta}|) / (\sum_{\boldsymbol{\kappa} \in \overline{\Omega}^{[1,30],P_\alpha}} \kappa_6 / |\overline{\Omega}^{[1,30],P_\alpha}|) \tag{4.18}$$

4.2.3.5 Experimentation

> 'Because of an observed association some preventative action is taken,
> Does it in fact prevent?' - **Bradford-Hill** [19].

The final Bradford-Hill causality consideration investigated is experimentation. There is deviation between the meaning behind experimentation, some authors assume it relates to intervention (i.e. if the drug stops does the medical event, if the drug restarts does the medical event follow?) [175], whereas others believe it corresponds literally to experiments that have been conduced and their results [64].

In this thesis we adopt the intervention interpretation and apply a retrospective investigation to find instances where a patient stopped taking the drug for a while and then restarted, and refer to this as a retrospective intervention. Unfor-

tunately, few patients experience retrospective interventions and this limits the experimentation attribute's usefulness. If a patient has two or more independent prescriptions of the drug, it can be observed whether the medical event often occurs shortly after the prescriptions but never shortly before. If this is the case, then this is a very strong implication that the medical event is an ADR.

Equation 4.19 shows the calculation for the proposed novel attributes based on the Bradford-Hill experimentation causality consideration. It is determined by finding the number of patients that have READ code β within 1 to 30 days of two or more independent prescriptions of α but never within 1 to 30 days before any prescriptions of α divided by the number of patients that have two or more independent prescriptions of α.

$$x_{26}^{\alpha\beta} = \frac{|\{\kappa_1 | \exists \boldsymbol{\kappa}, \overset{*}{\boldsymbol{\kappa}} \in \overline{\Omega}^{[1,30],P_\alpha,E_\beta}, \kappa_4 \neq \overset{*}{\kappa}_4, \kappa_1 = \overset{*}{\kappa}_1\} \bigcap \{\kappa_1 | \boldsymbol{\kappa} \notin \overline{\Omega}^{[-30,-1],P_\alpha,E_\beta}\}|}{|\{\kappa_1 | \exists \boldsymbol{\kappa}, \overset{*}{\boldsymbol{\kappa}} \in \overline{\Omega}^{[1,30],P_\alpha}, \kappa_4 \neq \overset{*}{\kappa}_4, \kappa_1 = \overset{*}{\kappa}_1\}|} \tag{4.19}$$

4.2.3.6 Other Criteria

The consistency, plausibility and coherence require additional resources for their calculation and are not tackled in this thesis. The analogy factor is indirectly incorporated by using a supervised algorithm, as attributes similarities for the drug-READ code pairs corresponding to known ADRs are learned and used to infer new ADRs.

4.2.3.7 THIN Specific

The attributes specific to the THIN database make use of the READ code structure and additional information available that might help distinguish between an

ADR and non-ADR. The first attribute gives a measure of how much noise there is present for the READ code β and drug α as it is harder to classify drug-READ code pairs with a large measure of noise as the attribute values are likely to be misleading. To determine the level of noise, the average number of prescriptions that occur within the two month interval centred around the prescription of α is calculated for all patients and compared with the average number of prescriptions that occur within the two month time interval centred around the prescription of drug α for the patients that also experienced READ code β within 1 to 30 days.

$$x_{27}^{\alpha\beta} = \Big(\sum_{\boldsymbol{\kappa} \in \overline{\Omega}^{[-30,30],P_\alpha E_\beta}} \kappa_8 / |\overline{\Omega}^{[-30,30],P_\alpha E_\beta}| \Big) / \Big(\sum_{\boldsymbol{\kappa} \in \overline{\Omega}^{[-30,30],P_\alpha}} \kappa_8 / |\overline{\Omega}^{[-30,30],P_\alpha}| \Big) \quad (4.20)$$

The next attribute investigates whether, for each READ code β that occurs within 1 and 30 days of a prescription of α, the patient has previously experienced β (or its level 3 parent). If many patients have previously experienced a more general version of β but not β itself, then this is a sign that β is not an ADR to drug α. The reason is that β is likely to have occurred due to an illness progression rather than it being caused by the drug. This prompts,

$$x_{28}^{\alpha\beta} = \Big(\sum_{\boldsymbol{\kappa} \in \Omega^{[1,30],P_\alpha E_\beta}} \kappa_{12} \Big) / \Big(\sum_{\boldsymbol{\kappa} \in \Omega^{[1,30],P_\alpha,E_\beta}} \kappa_{13} \Big) \quad (4.21)$$

The final THIN specific attributes use the READ code structure to help distinguish between associations that are causal and associations that are due to illness progressions. These attributes calculate the AB ratio when only considering the more general versions of all the READ codes. Therefore, if the association has occurred due to the illness progression, the AB ratio for the more general versions

of the READ codes should be small, even if the AB ratio for the actual READ code is large.

Letting Φ_3^β denote the set of READ codes that have the same level 3 parent as β, then the AB ratio is calculated on the transformed data,

$$x_{29}^{\alpha\beta} = |\bigcup_{\overset{*}{\beta}\in\Phi_3^\beta} \Omega^{[1,30],P_\alpha,E_{\overset{*}{\beta}}}|/|\bigcup_{\overset{*}{\beta}\in\Phi_3^\beta} \Omega^{[-30,-1],P_\alpha,E_{\overset{*}{\beta}}}| \tag{4.22}$$

Similarly, Φ_4^β denotes the set of READ codes that have the same level 4 parent as β and the AB ratio is calculated,

$$x_{30}^{\alpha\beta} = |\bigcup_{\overset{*}{\beta}\in\Phi_4^\beta} \Omega^{[1,30],P_\alpha,E_{\overset{*}{\beta}}}|/|\bigcup_{\overset{*}{\beta}\in\Phi_4^\beta} \Omega^{[-30,-1],P_\alpha,E_{\overset{*}{\beta}}}| \tag{4.23}$$

4.2.3.8 A Note on Dependency

Many of the attributes derived from the same Bradford-Hill causality considerations may have some statistical dependency but also give slightly different perspectives. The statistical dependency is unlikely to have any negative consequences on the future classifiers as either feature selection is applied to remove any redundancy or the methods are unaffected by statistical dependency. The random forest is a decision tree based classifier; decision trees partition the attribute space based on measures such as entropy. At each iteration the decision tree will simply pick the partitioning of an attribute space based on how well it separates the classes, dependency of two attributes will not have any negative effect on this process. For the other classifier used throughout this research, feature selection will be performed prior to training. Feature selection will choose a

subset of attributes to be used by the classifier that maximises its performance. If two dependant attributes negatively affect the classifier then the optimal feature subset will only contain a maximum of one of them.

4.2.4 Data Description

The attribute vector for a drug-READ code pair (α, β) is $\hat{\mathbf{x}}^{\alpha\beta} = (x_1^{\alpha\beta}, x_2^{\alpha\beta}, ..., x_{30}^{\alpha\beta}) \in \mathbb{R}^{30}$. The set $\hat{X}^{\alpha} = \{\hat{\mathbf{x}}^{\alpha\beta_i}, \beta_i \in RME^{\alpha}\}$ contains all the (α, β_i) corresponding Bradford-Hill causality consideration derived attribute vectors for the drug α and each of its 'risk' READ codes β_i (the READ codes recorded within 1 to 30 days from any prescription of α).

4.2.5 Data Transformation

4.2.5.1 Continuous Attributes

The importance of processing the data has been stressed in [127], one vital stage in processing is to ensure each attribute is treated equally by a classifier. This is done by normalising the data. Normalisation scales the nominal attribute data between two values [96], this ensures the optimal performance of some learning algorithms [127]. The three frequently implemented normalisation techniques are,

N1 (z-score Normalisation, useful if data bounds are unknown)

$$f : X \to X; f_{z-score}(x) = (x - \bar{X})/(\sigma_X)$$

N2 (Min Max Normalisation, data are scale into the range $[0, 1]$)

$$f : X \to X; f_{MinMax}(x) = (x - min_X)/(max_X - min_X)$$

Table 4.2: The results of KNN with leave one out cross validation when the different normalisations are applied to the data, k=8 (preliminary results showed this was optimal).

Normalisation	TP	FP	FN	TN	AUC
None	684	443	1080	7746	0.7885
N1	674	246	1090	7943	0.8055
N2	698	387	1066	7802	0.7567
N3	673	349	1091	7840	0.7561

N3 (Decimal Scaling Normalisation, data are scale into the range $[-1, 1]$)

$$f : X \to X; f_{decimal}(x) = x/10^j, j = \min\{j \in \mathbb{N} | \forall_{x_i \in X} |x_i/10^j| < 1\}$$

Table 4.2 shows the results when a KNN algorithm was applied with leave one out cross validation on labelled data for 25 drugs with the data transformations N1-N3 and no transformation. The optimal solution was obtained when the N1 transformation was applied, as the AUC was the greatest. Therefore, the N1 transformation will be used to transform any continuous attributes prior to any learning algorithm in the remainder of this thesis.

4.2.5.2 Discrete Attributes

The discrete attributes, $x_{15}^{\alpha\beta} - x_{18}^{\alpha\beta}$ are not normalised. The binary attributes do not require any transformation, but the non-binary discrete attribute $x_{18}^{\alpha\beta}$ corresponding to the READ code hierarchal level does. As described in [127], dummy attributes (binary attributes corresponding to the each value of the discrete attribute) are generated for $x_{18}^{\alpha\beta}$, see Table 4.3 for an example. Due to the linear dependancy between $x_{18}^{\alpha\beta}$'s dummy attributes, one can be discarded, so only four of the dummy attributes are used. If we denote the z-score normalisation of the continuous attributes and the creation of dummy attributes for the discrete at-

Table 4.3: Example of how a discrete attribute is transformed into its dummy attributes.

Orignal	Dummy Attributes				
READ code lv	lv1	lv2	lv3	lv4	lv5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
3	0	0	1	0	0

tributes by the mapping $f : \mathbb{R}^{30} \rightarrow \mathbb{R}^{33}$, then this leads to the transformed data $X^{\alpha} = \{f(\mathbf{x}), x \in \hat{X}^{\alpha}\}$.

4.2.6 Feature Selection

In preliminary work I investigated the usefulness of different Bradford-Hill causality considerations based attributes for predicting ADRs. A multivariate filter known as the Correlation-based Feature Selection (CFS) algorithm [69] was applied to a range of attributes. This feature selection technique aims to find a subset of attributes such that each individual attribute in the subset is more correlated to the label/class than it is to other attributes in the subset. Therefore, only attributes that offer new insight for predicting the label/class are included, the others will be removed, as they are generally redundant. The paper detailing this preliminary work can be found in appendix C. However, in the actual framework proposed within this work, I will use wrapper feature selection prior to the classification (except for random forest) as wrapper feature selection chooses the attribute subset based on how well the classifier performs when trained using only the subset of attributes, this is more useful than a multivariate filter as it

considers the classifier performance rather than just relying on correlations.

A wrapper approach to feature selection implements a heuristic search through the power set of the attributes. It aims to find the attribute subset that, when used as input into a classifier, maximises the classifiers performance. For example, consider the attribute set $\{x_1, x_2, x_3\}$, the power set of these attributes is $\{\{\}\{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}\}$. A wrapper that performs an exhaustive search will apply the classification using every possible subset of attributes (all the power set) and choose the subset for which the classifier performed the best.

In general it is not suitable to perform an exhaustive search and a local optimal subset (hopefully with a good performance) will be found instead. A forward search starts with one attribute and iteratively investigates the addition of a single attribute at a time until there is no further improvement possible. A backwards search starts with all the attributes and iteratively investigates the removal of a single attribute at a time until there is no further improvement possible. These searchers described above are referred to as 'greedy' as the process of adding (or removing) an attribute cannot be reversed once done. This leads to the searchers finding local optimal subsets rather than global optimal ones.

In the future work I will implement a greedy backwards feature selection algorithm named 'rfe' available in the R caret package. This algorithm requires inputting the size of the attribute subset desired. It iteratively removes attributes based on their ranking of importance by the naive Bayes classifier until the attribute subset is reduced to the desired size. I will search for the subsets of size 5, 10, 15, 20, 25, 30 and 33. I will then select the subset out of these seven that maximises the classifier performance (prediction accuracy). Naive Bayes was chosen

as this classifier assumes conditional independence, so it is more likely to be negatively affected by attributes' dependencies. Details about the chosen attributes returned by wrapper feature selection throughout this research can be found in appendix C.2.

4.3 Summary

In this chapter methods to generate thirty-three attributes for each drug-READ code pair have been proposed. These attributes may help identify causal relationships as they are derived from the Bradford-Hill causality considerations [19] that has been used frequently to investigate causality between a single drug-medical event pair. In addition, THIN specific attributes were presented with the aim of preventing issues that arise due to the hierarchal READ code structure. The attributes were explored and it was determined that z-score normalisation should be applied to transform the continuous data and optimise the results of any learning algorithms applied. The foundations have been set to enable the development of a novel framework for ADR signalling that incorporates causal knowledge and the THIN data structure.

In Chapter 5, the attributes proposed in this chapter and their transformations will be used as inputs into a learning algorithm which will be trained, using the knowledge of known ADRs and non-ADRs, to distinguish between causal and non-causal relationships.

Chapter 5

Developing The ADR Learning Framework

> 'Our evaluation showed that the phenotypic information (when available) largely improved the performance of ADR prediction models. '
>
> M. Liu [113]

5.1 Introduction

In the previous chapter suitable attributes based on the Bradford-Hill causality considerations and specific to the THIN database were proposed, with the aim of being used as inputs into a learning algorithm capable of identifying causality and hence, able to signal ADRs. This was the first step toward testing the main hypothesis being investigated in this thesis, that a framework that incorporates attributes that give insight into causality and attributes specific to the THIN database into a learning algorithm that uses knowledge of existing ADRs will

signal ADRs without a high false positive rate. The second step is determining which learning algorithm is optimal.

In this chapter the focus is on developing an algorithm for detecting ADRs by applying supervised and semi-supervised techniques that utilised knowledge of existing ADRs and non-ADRs. The sub question answered in this chapter is what will yield a better ADR signalling algorithm, a supervised approach that is trained on labelled data corresponding to a variety of drugs, or a semi-supervised approach that uses the labelled and non-labelled data for a single drug?

5.2 Motivation

Generally ADR signalling methods using data contained in SRS databases and LODs have been unsupervised, however, numerous supervised algorithms have been presented to classify ADR using chemical structures and known ADRs. One of the first algorithms that used chemical structures to infer ADRs was developed for a specific group of drugs known as the CEPT inhibitors [209]. This idea was expanded to simultaneously identify multiple ADRs [5] where the authors prosed two novel algorithms that incorporate knowledge of chemical structures and known ADRs, extracted from SIDER [98], to infer new ADRs. The first algorithm learns associations between drug attributes and known ADRs and uses this knowledge to infer new ADRs. The second algorithm, based on a method of predicting disease-causing genes [186], uses a diffusion process that incorporates the similarities between drugs and the similarities between ADRs. The overall measure of how likely a drug causes an ADR was calculated using a combination of the values returned by the two algorithms. More recent methods have utilised target

protein information in addition to chemical structure and ADR knowledge, and used these to generate attributes that are fed into a predictive model [109; 212] or included biological and phenotypic (e.g., indications and known ADRs) based attributes [113]. In [113] the presented framework detected ADRs with a precision of 66.17% and a recall of 63.06% and it was shown that including attributes based on known ADRs improves the ability of the classifier and increases the recall and precision. This is a key result, as it shows that incorporating knowledge of known ADRs into an algorithm for signalling ADRs may decrease the false positive rate.

ADR signalling algorithms applied to LODs have exceptional potential to identify new ADRs [204], but are currently limited by the high number of false positives [156]. An ADR signalling algorithm that generates attributes based on LODs, but also incorporates known ADR labels may reduce the number of false positives and should outperform the existing unsupervised algorithms. As an ADR represents a causal relationship, any attributes used by a learning algorithm to distinguish between ADRs and non-ADRs need to contain information about causality. In Chapter 4 suitable attributes for each drug-medical event pair based on the strength, temporality, specificity, biological gradient and experimentation factors of the Bradford-Hill causality considerations were investigated, as well as attributes specific the to THIN database. These attributes are suitable inputs into a causality learning algorithm as they are frequently implemented by researchers to determine causal relations [42]. In Chapter 3.4, labels were extracted from the SIDER resource for known ADRs, and noise medical events were extracted using the READ code tree. Using the generated attributes and the known labels it may be possible to learn areas of the attribute space that suggest a drug-medical event pair represent an ADR. It is hypothesised that such an algorithm would

reduce the time required to definitively identify ADRs and enable a wider search for ADRs. Overall, this would improve current healthcare.

5.3 Algorithms

After the THIN data has been processed, as described in Chapter 4, for each drug of interest α, we can assign class labels to some of the drug-READ code pairs (class ADRs or class non-ADRs). Therefore, the set of Bradford-Hill causality consideration attribute vectors X^α is partitioned into labelled data, $(X_L^\alpha, Y_L^\alpha) = \{(\mathbf{x}, y)|\mathbf{x} \in X^\alpha \text{ and } y \text{ is the known label}\}$, and unlabelled data, $X_U^\alpha = \{(\mathbf{x}, y)|\mathbf{x} \in X^\alpha \text{ and the label is unknown}\}$. The aim of the learning algorithms is to determine a predictive function $f : X \to Y$, using the labelled data and the unlabelled data, that can then be applied to the attributes of a new drug-medical event pair, (α^*, β^*), to predict the pair's class.

5.3.1 Supervised ADR Predictor

Supervised methods only use the labelled data (X_L^α, Y_L^α). The Supervised ADR Predictor (SAP) framework signals ADRs by applying a classifier that is trained on n drugs to a drug not used to train the classifier. Using a sufficiently large value for n ensures that there is an adequate number of labels. In this study the value of n used is 24 as this corresponded to approximately $10,000$ labelled data-points. As the drug being investigated is not used in the training, no knowledge of existing ADRs for that drug is required, so the SAP framework can be applied to newly marketed drugs. The framework is illustrated in Figure 5.1.

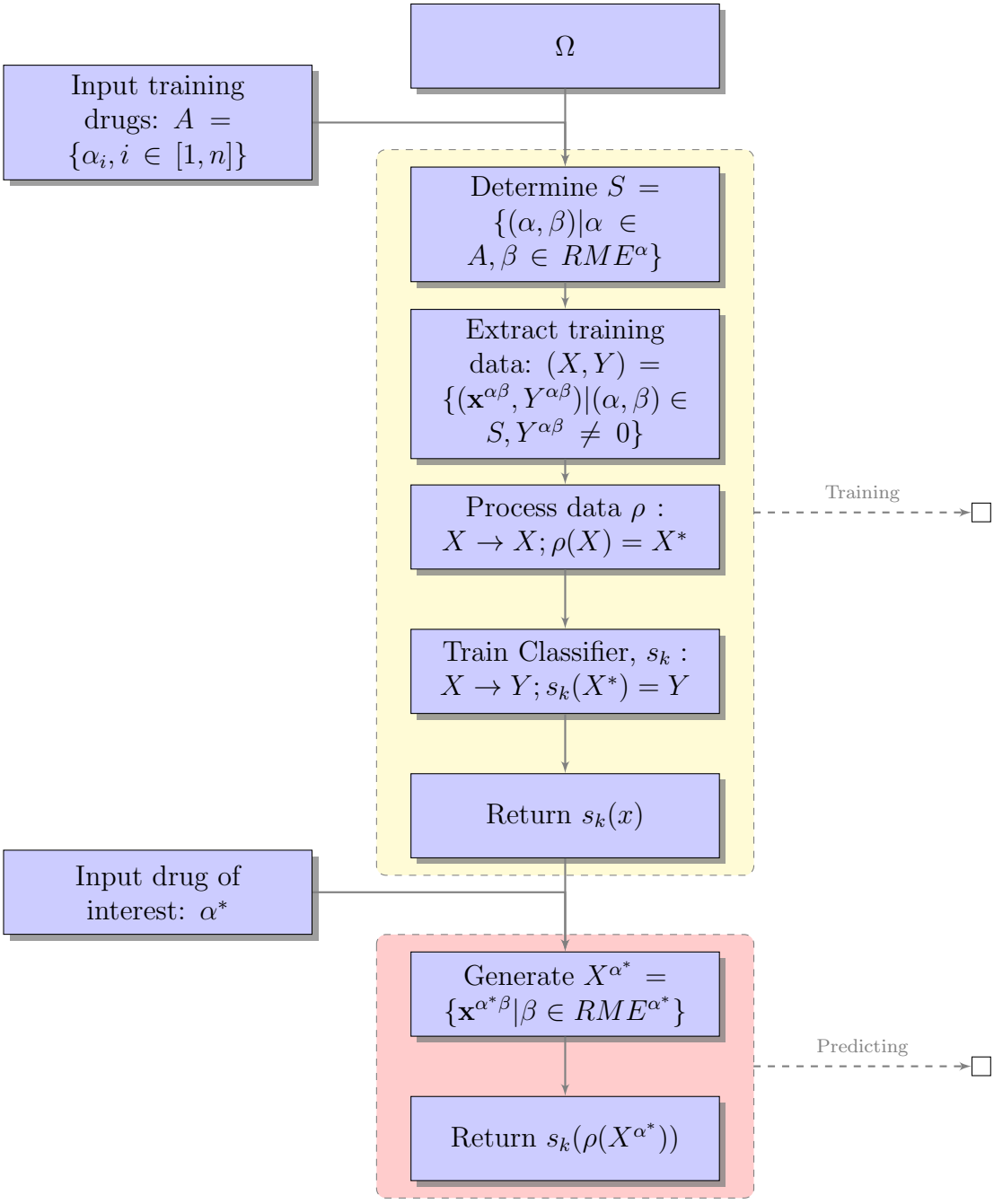


Figure 5.1: The framework implemented to train the four different classifiers using a variety of n drugs with known side effects. These general classifiers are then used to predict the class for unlabelled data.

5.3.1.1 Training Stage

Starting with the THIN data, Ω, and the set of training drugs, $A = \{\alpha_i, i \in [1, 24]\}$, the first step is determining the 'risk' drug-medical event pairs for each of the training drugs. These are all the medical event and drugs pairs,(α_i, β) , where the medical event β is observed to be recorded, for at least one patient, within 30 days after a training drug α_i was recorded. The set containing all these 'risk' drug-medical event pairs where the drug is a training set drug is denoted by S.

Next, the Bradford-Hill causality consideration based attributes and the THIN specific attributes are extracted for each pair in S with a corresponding label that is ± 1. The labels are determined (using the sets Ψ^A and Ψ^N defined in Chapter 3.4) by,

$$Y^{\alpha\beta} = \begin{cases} 1 & \text{if } (\alpha, \beta) \in \Psi^A; \\ -1 & \text{if } (\alpha, \beta) \in \Psi^N. \\ 0 & \text{else.} \end{cases} \tag{5.1}$$

So, the extracted labelled data is, $(X, Y) = \{(\mathbf{x}^{\alpha\beta}, Y^{\alpha\beta}) | (\alpha, \beta) \in S, Y^{\alpha\beta} \neq 0\}$. Before the labelled data is used to train the classifier, it is processed according to the chosen classifier being implemented. The processing step is represented by the function $\rho : X \to X$. For the random forest classifier, ρ is the z-score normalisation function, $\rho(\mathbf{x}) = (\mathbf{x} - \mu)/\sigma$, where μ is the mean of X and σ is the standard deviation of X. For the SVM, Naive Bayes and Logistic Regression, ρ represents z-score normalisation and wrapper feature selection (see appendix C.2) [157].

The final step is using the processed labelled data to train and return the clas-

sifier, $s_k : X \rightarrow Y$, using leave one out cross-validation. The classifier is trained so that $s_k(\mathbf{x}^{ab}) = 1$ represents drug a causing medical event b and $s_k(\mathbf{x}^{ab}) = -1$ represents drug a not causing medical event b. The trained random forest classifier is represented by $s_1 : X \rightarrow Y$, the SVM is represented by $s_2 : X \rightarrow Y$, the Logistic regression is represented by $s_3 : X \rightarrow Y$ and the Naive Bayes classifier is represented by $s_4 : X \rightarrow Y$.

5.3.1.2 Prediction Stage

Once the classifier, $s_k : X \rightarrow Y$, is trained, the SAP algorithm can then be applied to any drug α^* not used in training. The set of attribute vectors, $X^{\alpha^*} = \{\mathbf{x}^{\alpha^*\beta} | \beta \in RME^{\alpha^*}\}$, corresponding to 'risk' drug-medical event pairs containing the drug being investigated are extracted and processed. The trained classifier is then applied to each of the data-points, $\mathbf{x}^{\alpha^*\beta} \in X^{\alpha^*}$, to predict whether the drug α^* and medical event β correspond to an ADR. The final output for classifier k is the set of medical events that correspond to the signalled drug-medical event pairs containing drug α^*, $\{\beta | s_k(\mathbf{x}^{\alpha^*\beta}) = 1, \mathbf{x}^{\alpha^*\beta} \in X^{\alpha^*}\}$.

5.3.1.3 Results and Analysis

To analyse the SAP algorithm a set of 25 drugs were chosen, $D = \{\alpha_i, i \in [1, 25]\}$. For each drug, $\alpha_i \in D$, the SAP algorithm was trained on the set of drugs in D excluding α_i and then validated by being applied to α_i. The inputs into the SAP framework were, $A = \{\alpha_j \in D | j \neq i\}$ and $\alpha^* = \alpha_i$. The predictions of the classifiers on each labelled data-point (not used during training), $s_k(\mathbf{x}^{\alpha_i\beta})$, are then compared with the truth, $Y^{\alpha_i\beta}$, to measure the ability of the classifier. Using the validation set, $(X^{\alpha_i}, Y^{\alpha_i}) = \{(\mathbf{x}^{\alpha_i\beta}, Y^{\alpha_i\beta}) | \beta \in RME^{\alpha_i}, Y^{\alpha_i\beta} \neq 0\}$, the

Table 5.1: The results of the different classifiers at their natural thresholds for three drugs, Nifedipine, Ciprofloxacin and Ibuprofen.

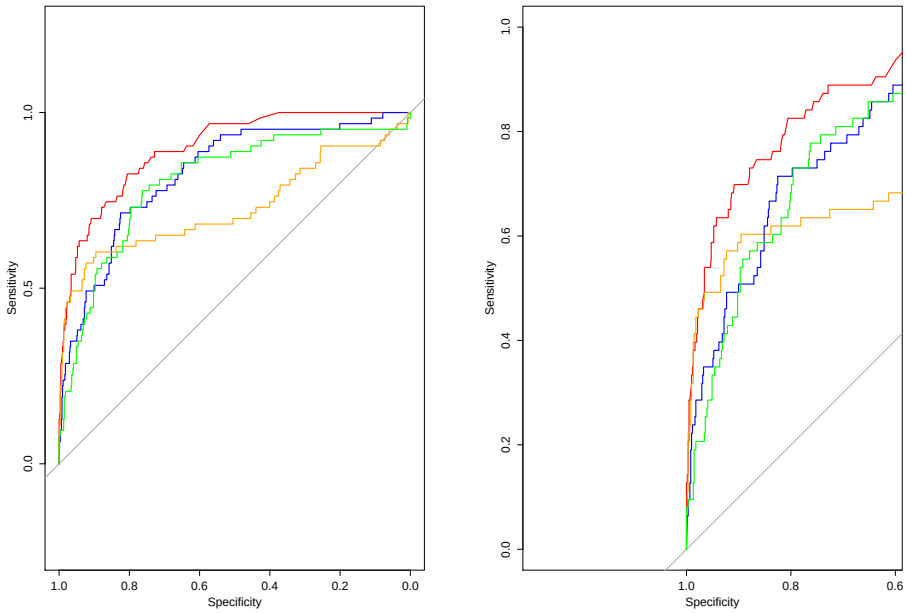
Drug	Algorithm	TP	FN	FP	TN	Sensitivity	Specificity
Nifedipine	RF	22	41	8	604	0.349	0.987
Ciprofloxacin	RF	13	42	0	385	0.236	1
Ibuprofen	RF	22	53	2	784	0.293	0.997
Nifedipine	NB	14	49	6	606	0.222	0.990
Ciprofloxacin	NB	6	49	5	380	0.109	0.987
Ibuprofen	NB	9	66	16	770	0.120	0.980
Nifedipine	SVM	25	38	9	603	0.397	0.985
Ciprofloxacin	SVM	12	43	0	385	0.218	1
Ibuprofen	SVM	24	51	9	777	0.32	0.989
Nifedipine	LR	6	57	2	610	0.095	0.997
Ciprofloxacin	LR	6	52	6	379	0.103	0.984
Ibuprofen	LR	12	63	7	779	0.16	0.991

number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), for each classifier s_k, are calculate as,

- $TP = |\{s_k(\mathbf{x_j}) = y_j | (\mathbf{x_j}, y_j) \in (X^{\alpha_i}, Y^{\alpha_i}), y_i = 1\}|$
- $FP = |\{s_k(\mathbf{x_j}) = 1 | (\mathbf{x_j}, y_j) \in (X^{\alpha_i}, Y^{\alpha_i}), y_i = -1\}|$
- $FN = |\{s_k(\mathbf{x_j}) = -1 | (\mathbf{x_j}, y_j) \in (X^{\alpha_i}, Y^{\alpha_i}), y_i = 1\}|$
- $TN = |\{s_k(\mathbf{x_j}) = y_j | (\mathbf{x_j}, y_j) \in (X^{\alpha_i}, Y^{\alpha_i}), y_i = -1\}|$

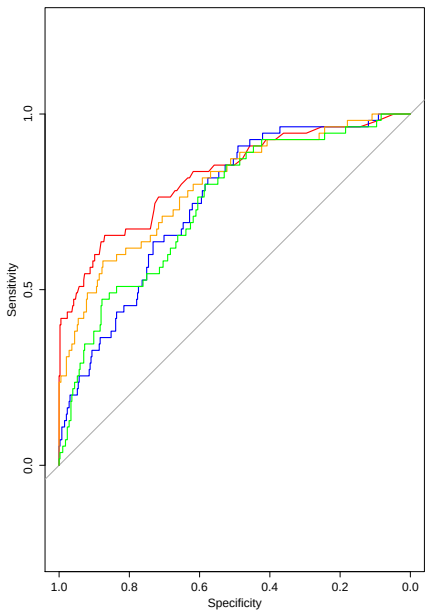
Table 5.1 presents the results of the classifiers at their natural threshold and Figure 5.2 presents the ROC plots and partial ROC plots of the classifiers, respectively.

Bar charts of the AUC and $AUC_{[0.9,1]}$ values returned by the classifier for the three drugs are displayed in Figure 5.3. The random forest classifier had significantly greater $AUC_{[0.9,1]}$s for all three drugs investigated at a 5% significance level. However, the random forest's AUC was only significantly greater for

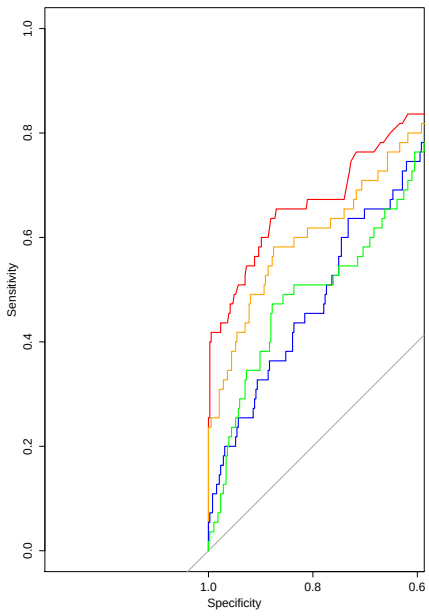


(a) Nifedipine: all specificity

(b) Nifedipine: high specificity

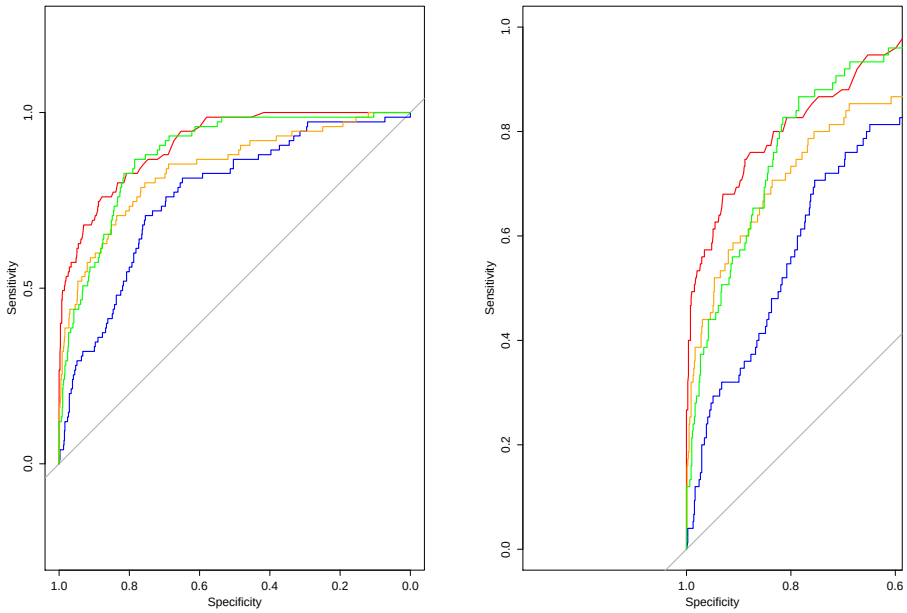


(c) Ciprofloxcin: all specificity



(d) Ciprofloxcin: high specificity

Figure 5.2: The ROC curves for the different classifiers used to predict the ADRs of the drugs. The red curve represents the random forest classifier, the orange curve represents the support vector machine classifier, the green curve represents the logistic regression classifier and the blue curve represents the Naive Bayes classifier.



(e) Ibuprofen: all specificity (f) Ibuprofen: high specificity

Figure 5.2: Continuation of the ROC plots.

Nifedipine (p-value: 0. 0002) and not for Ciprofloxacin nor Ibuprofen (p-values 0.164 and 0.052 respectively). The AUC for the random forest classifier was in the range $[0.823, 0.912]$, indicating excellent performance. The other classifiers also performed well with the lowest AUC value of 0.730 obtained by the SVM for Nifedipine.

The random forest and SMV were able to signal between $24\% - 35\%$ and $22\% - 40\%$ of the known ADR READ codes for the three drugs respectively. Interestingly, the classifiers all managed to keep the number of false positives low, aggregating over the three drugs, 85%, 77%, 52% and 62% of the signals returned by the random forest, SVM, Naive Bayes and logistic regression classifiers, respectively, were known ADRs. Consequently, although only approximately 30% of known ADRs were signalled by the random forest classifier at its natural thresh-

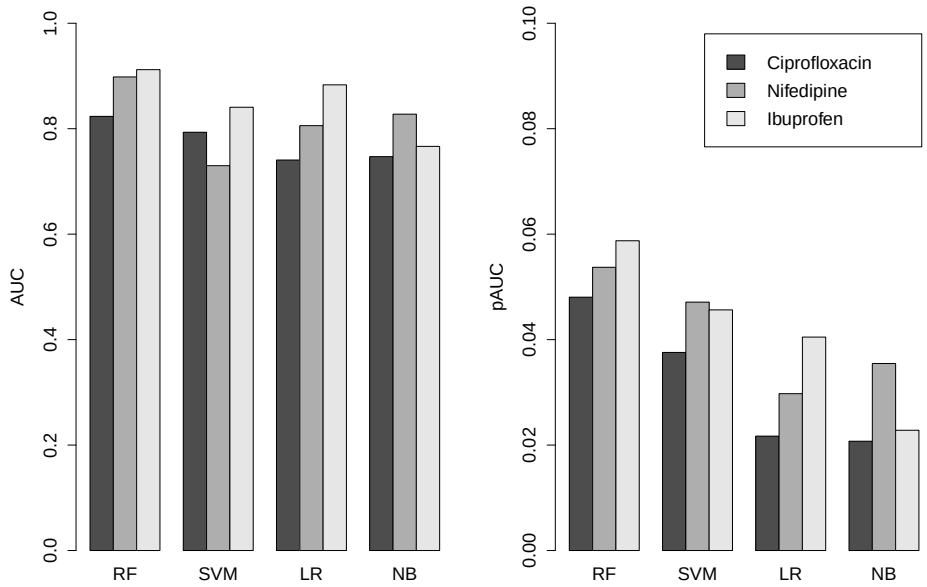


Figure 5.3: The AUC and $AUC_{[0.9,1]}$ values for the SAP algorithm implementing each of the classifiers when applied for the drugs Nifedipine, Ciprofloxacin and Ibuprofen.

old, the majority of signals were true, so no additional filtering would have been required to further validate the signals. This is an improvement over the existing methods (see Chapter 3).

When investigating the signal overlap between the classifiers, in general the SVM returned the greatest number of unique signals (6 for Ibuprofen, 6 for Nifedipine and 2 for Ciprofloxacin) although the random forest also had some unique ones (4 for Ibuprofen, 2 for Nifedipine and 4 for Ciprofloxacin). This suggests it may be of interest to investigate applying an ensemble technique that integrates the results obtained from all four classifiers.

5.3.1.4 Summary

The classifiers that use the Bradford-Hill causality consideration based and THIN specific attributes and additional knowledge of known ADRs and non-ADRs show excellent promise at effectively signalling ADRs. These classifiers are trained on drugs that are not investigated, so there is no requirement of known ADRs for the drugs investigated and the classifiers had a high specificity and sensitivity. Out of the four classifier investigated, the random forest returned significantly better results and was also the classifier that required the least amount of pre-processing, making it highly efficient. All four classifiers obtained a sufficiently high specificity in addition to constraining the number of false positives. The benchmark AUC for the supervised classifiers is set at 0.91 and the benchmark $AUC_{[0.9,1]}$ are set at 0.048 for Ciprofloxacin, 0.059 for Ibuprofen and 0.053 for Nifedipine.

5.3.2 Semi-Supervised ADR Predictor

In the previous section a general classifier was proposed but the combinations of attribute values that suggest a drug-medical event pair corresponds to an ADR may vary for each drug, so information may be lost by combining the labelled data for a variety of drugs. However, it is difficult to determine a specific classifier with a high accuracy as the number of known ADRs per drug is generally less than a hundred or so, but the number of 'risk' drug-medical event pairs are often in the thousands. When there is only a small number of labelled data, but surplus unlabelled data, it has been shown that semi-supervised techniques may yield more accurate results [126]. In this section a novel semi-supervised framework is proposed.

A frequently implemented example of a semi-supervised technique is the self-training wrapper algorithm, summarised in Chapter 2.2.2.1. This algorithm trains a classifier on labelled data and then gets the classifier to 'teach' itself by applying the trained classifier on the unlabelled data and adding any unlabelled data-point and its prediction to the labelled data, if the classifier is confident of the prediction [215]. In [195], the authors showed that the performance of a self-training approach and supervised approach is comparable by applying self-training using a tree based classifier to a natural language classification problem. This motivates the investigation of a self-training approach that incorporates a random forest to learn from the labelled and unlabelled data. Unfortunately, the self-training approach requires a sufficient number of initial labelled data, as classifiers perform poorly when trained on a small set of data [55], and an incorrect initial model will get interactively worse. When the size of the initial labelled data is

small, a semi-supervised clustering algorithm may be more appropriate as a small number of labels can be used to aid clustering by adding bias [7]. Therefore, in this Section, two semi-supervised algorithms are presented to signal ADRs using labelled and unlabelled data corresponding to one drug; a self-training random forest and a semi-supervised k-means clustering.

The framework for Semi-Supervised ADR Predictor (SSAP), that contains both the self-training algorithm and the semi-supervised clustering, is presented in Figure 5.4. The value $crit_*$ represents the critical value that is used to determine whether the self-training or semi-supervised clustering is applied during the SSAP framework, based on the fraction of total data that is labelled. This value will be determined by investigating the performance of both semi-supervised techniques when applied to data with a range of labelled data sizes.

5.3.2.1 Self Training Random Forest

The self training random forest iteratively trains a random forest on the labelled data for drug α, $(X_L^\alpha, Y^\alpha) = \{(\mathbf{x}^{\alpha\beta}, Y^{\alpha_i\beta})|\beta \in RME^\alpha, Y^{\alpha\beta} \neq 0\}$, but after each random forest is built, it is applied to the unlabelled data $X_U^\alpha = \{\mathbf{x}^{\alpha\beta}|\beta \in RME^\alpha, Y^{\alpha\beta} = 0\}$ and any unlabelled data point assigned a predicted class with a confidence greater than 0.9 is removed from the unlabelled set and added to the labelled set. The self training stops when the stopping criteria is met, either all the originally unlabelled data-points are moved into the labelled set or the iteration has run for twenty times. In detail, the self train process is:

Once the final random forest is trained, the final iteration model $\hat{s} : X \to Y$, is applied to the unlabelled data X_U^α. The algorithm returns the predicted class of the unlabelled data, $\hat{s}(\mathbf{x}), \mathbf{x} \in X_U^\alpha$, or the confidence of the data point being in

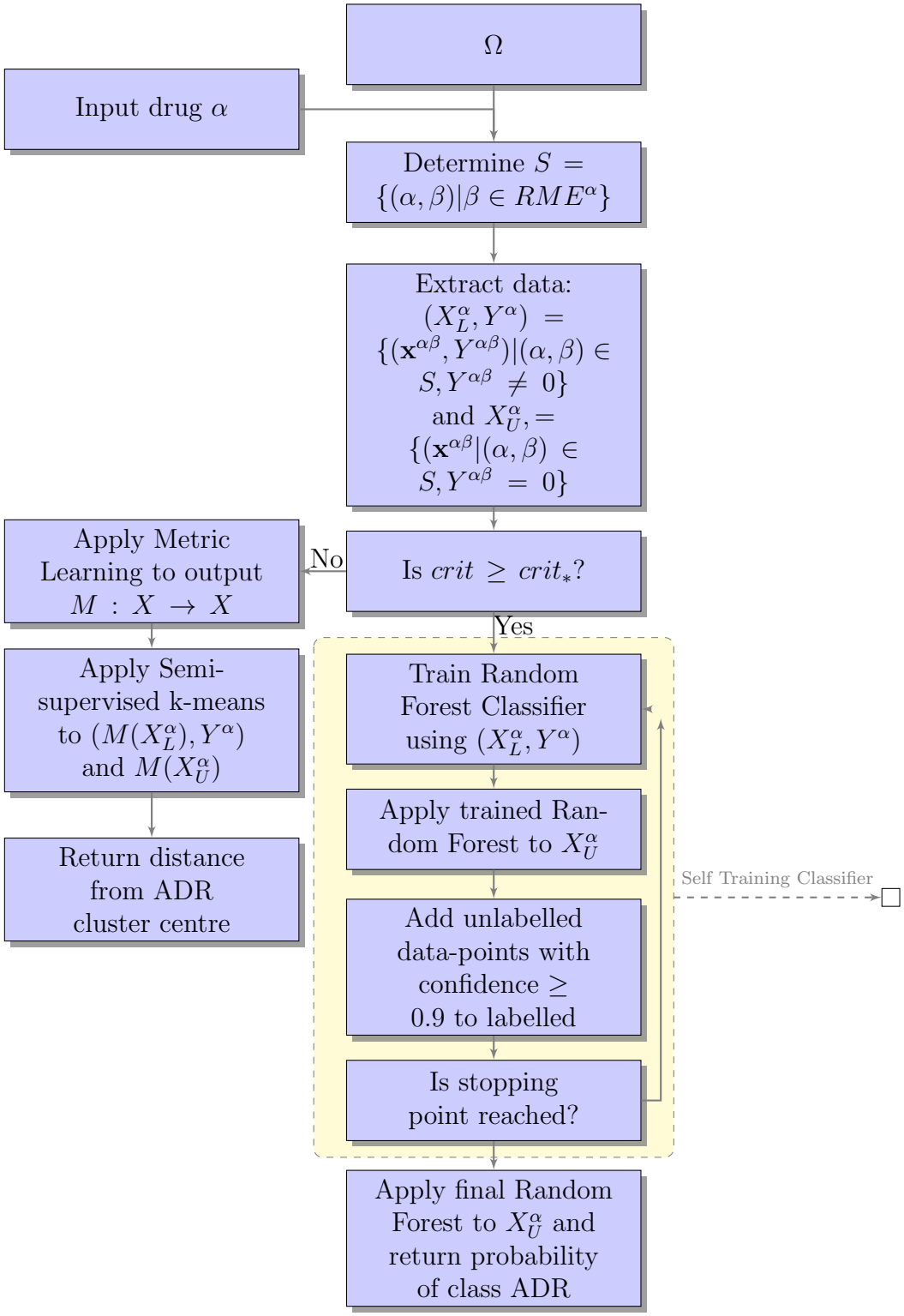


Figure 5.4: The framework for the Semi-Supervised ADR Predictor algorithm. This algorithm uses labelled and unlabelled data for the drug of interest only during training. The technique applied depends on the percentage of labelled data.

Input : Labelled data: (X_L^α, Y^α), Unlabelled data: X_U^α and Iteration limit: $n\}$
Output: Final random forest applied to unlabelled data, $s^i(X_U^\alpha)$
Initialization: $L^1 = \emptyset$, $U^1 = X_U^\alpha$
for $i = 1, 2, 3, ...$ **do**
  Train random forest, $s^i : X \to Y$, on labelled data $(X_L^\alpha, Y^\alpha) \bigcup L^i$.
  Apply to unlabelled data $s^i(U^i)$, Set:
  $L^{i+1} = L^i \bigcup \{(\mathbf{x}^{\alpha\beta}, s^i(\mathbf{x}^{\alpha\beta}))|\mathbf{x}^{\alpha\beta} \in U^i,$ confidence of prediction $\geq 0.9\}$
  $U^{i+1} = \{\mathbf{x}^{\alpha\beta}|\mathbf{x}^{\alpha\beta} \in U^i,$ confidence of prediction $< 0.9\}$
  if $i \geq n$ or $U^{i+1} = \emptyset$ then **break**
end

Algorithm 3: The self train random forest algorithm

the ADR class.

5.3.2.2 Semi-supervised Clustering

The semi-supervised clustering technique is proposed for replacing the self-training random forest when there is insufficient number of labelled data. The semi-supervised clustering method has two steps, the first step applies metric learning [211] using the labelled data, X_L^α, to learn a mapping, $M : X \to X$, of the attribute space that minimises the distance between data-points in the same class while adding a constraint to keep data-points from different classes sufficiently far apart. The second step is the application of the seed-constrained k-means clustering algorithm [7] to determine the clusters using the mapped data, $M(X^\alpha)$, where $X^\alpha = X_U^\alpha \cup X_L^\alpha$. The k-means algorithm uses the labelled data to determine the initial centres of the clusters and fixes the labelled data to a cluster, the unlabelled data-points are then iteratively assigned to the cluster with the closest centre until convergence. Both the metric learning and the semi-supervised k-means algorithm are described in Chapter 2.2.2.1. In detail, the process is:

1. Apply metric learning with $\mu = 10^{-5}, tol = 1-^{-5}$ and $\alpha_t = 0.02$, where:

   - $X = X_L^\alpha \bigcup X_U^\alpha$
   - $S = \{(i,j)|(\mathbf{x_i}, y_i), (\mathbf{x_j}, y_j) \in (X_L^\alpha, Y^\alpha), y_i = y_j\}$
   - $D = \{(i,j)|(\mathbf{x_i}, y_i), (\mathbf{x_j}, y_j) \in (X_L^\alpha, Y^\alpha), y_i \neq y_j\}$

   To output the metric space mapping $M : X \to X$

2. Apply seed-constrained k-means to $M(X)$, where:

   - $K = 2$
   - $S_1 = \{M(\mathbf{x_i})|(\mathbf{x_i}, y_i) \in (X_L^\alpha, Y^\alpha), y_i = 1\}$
   - $S_2 = \{M(\mathbf{x_i})|(\mathbf{x_i}, y_i) \in (X_L^\alpha, Y^\alpha), y_i = -1\}$

   To output the final cluster or the distance from the final ADR centre, $||M(\mathbf{x_i}) - \mu_\mathbf{1}||$, where $\mu_\mathbf{1}$ is the centre of the ADR cluster.

The algorithm returns the predicted cluster of the unlabelled data-points or the distance between the data point and the ADR cluster centre.

5.3.2.3 Results and Analysis

The self-training random forest algorithm and semi-supervised clustering algorithm implemented the SSAP framework were both applied to the drugs Nifedipine, Ciprofloxacin and Ibuprofen to analyse the results. For each drug, α_i, the labelled data was extracted, $(X^{\alpha_i}, Y^{\alpha_i}) = \{(\mathbf{x}^{\alpha_i\beta}, Y^{\alpha_i\beta})|\beta \in RME^{\alpha_i}, Y^{\alpha_i\beta} \neq 0\}$, and randomly partitioned into disjoint training, $(X_L^{\alpha_i}, Y_L^{\alpha_i})$, and validation, $(X_U^{\alpha_i}, Y_U^{\alpha_i})$, sets.

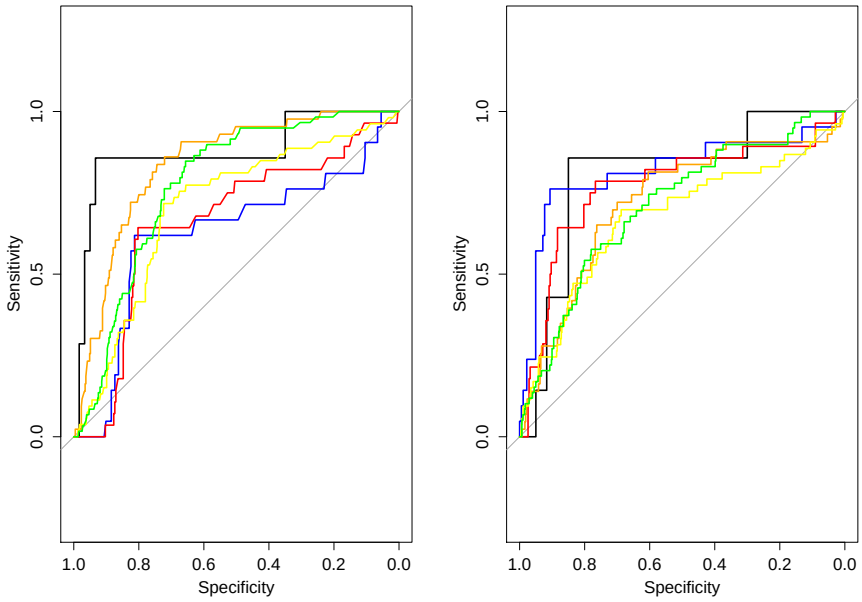
$$(X^{\alpha_i}, Y^{\alpha_i}) = (X_L^{\alpha_i}, Y_L^{\alpha_i}) \bigcup (X_U^{\alpha_i}, Y_U^{\alpha_i})$$

$$(X_L^{\alpha_i}, Y_L^{\alpha_i}) \bigcap (X_U^{\alpha_i}, Y_U^{\alpha_i}) = \emptyset$$

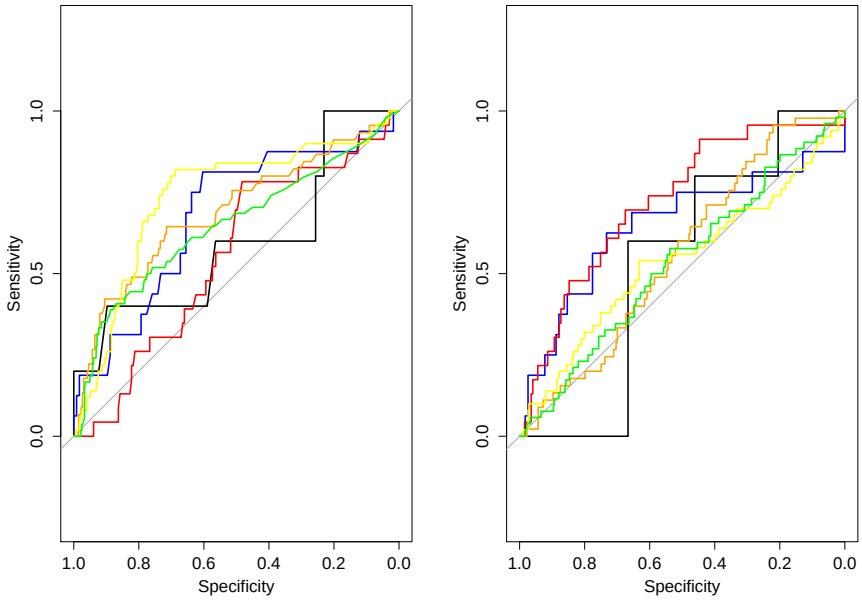
The data input into the algorithms as labelled data was $(X_L^{\alpha_i}, Y_L^{\alpha_i})$ and the data input into the algorithms as unlabelled data was $X_U^{\alpha_i}$. The algorithms were analysed by investigating the ROC plots determined by comparing their predictions on the validation data $\hat{s}(\mathbf{x}), \mathbf{x} \in X_U^{\alpha_i}$ with the truth $Y_U^{\alpha_i}$. Both algorithms were applied for varying values of $crit = |X_L^{\alpha_i}|/|X^{\alpha_i}|$, to investigate if there is an obvious threshold value of *crit* that can be used to determine which semi-supervised algorithm to apply when the SSAP framework is implemented (i.e., does the self-training algorithm always outperform the semi-supervised clustering when $crit \geq crit_*$?).

Figure 5.8 displays the AUC of the ROC plots obtained by applying the semi-supervised clustering or self-train classifier to the drugs Nifedipine, Ciprofloxacin, and Ibuprofen. The AUCs varied between $0.55-0.88$ and $0.53-0.84$ for the semi-supervised clustering and self-train classifier respectively. For Nifedipine, both algorithms performed their respective best, with AUCs of 0.80 and 0.88 for the clustering and self-train respectively, when 90% of the data was labelled, whereas for Ibuprofen, both algorithms performed their respective best when only 5% of the data was labelled. This shows that the semi-supervised techniques did not always improve in performance when the value of *crit* was increased, this is further evident in Figures 5.5-5.6. Futhermore, this suggests that the SSAP framework does not require a large number of labelled data, as, in general, the performance seems to be similar for low and high values of *crit*, but the performance depends on the quality of labelled data.

To investigate how much the initial labels affect the performance, both semi-



(a) Nifedipine -Left Plot: Self Training, Right Plot: Clustering



(b) Ciprofloxacin - Left Plot: Self Training, Right Plot: Clustering

Figure 5.5: The ROC curves for the SSAP framework at 6 different values of *crit* when applied to the different drugs. The black, blue, red, orange, yellow and green curve correspond to *crit* values of 0.9, 0.7,0.5,0.3,0.1 and 0.05 respectively.

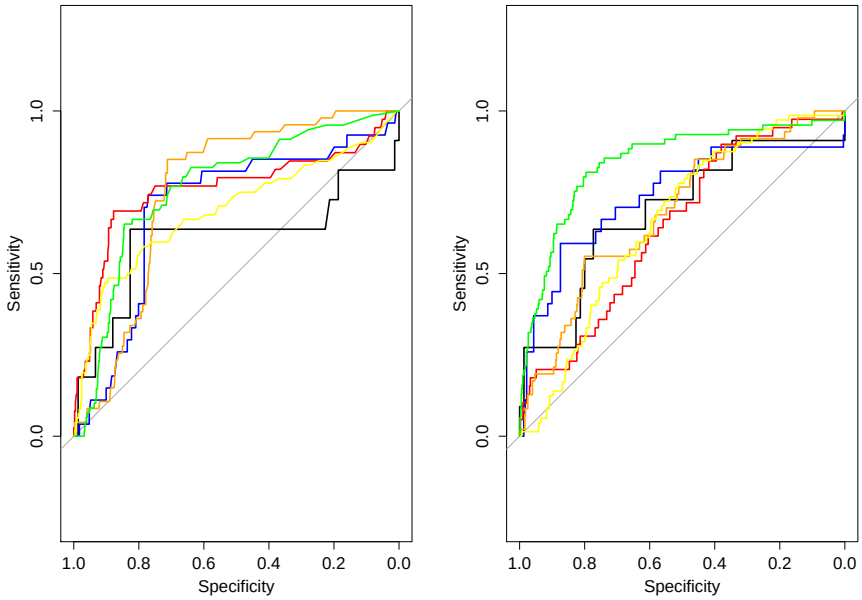


Figure 5.6: The ROC curves for the SSAP framework at 6 different values of *crit* when applied to Ibuprofen. The black, blue, red, orange, yellow and green curve correspond to *crit* values of 0.9, 0.7,0.5,0.3,0.1 and 0.05 respectively. Left Plot: Self Training, Right Plot: Clustering.

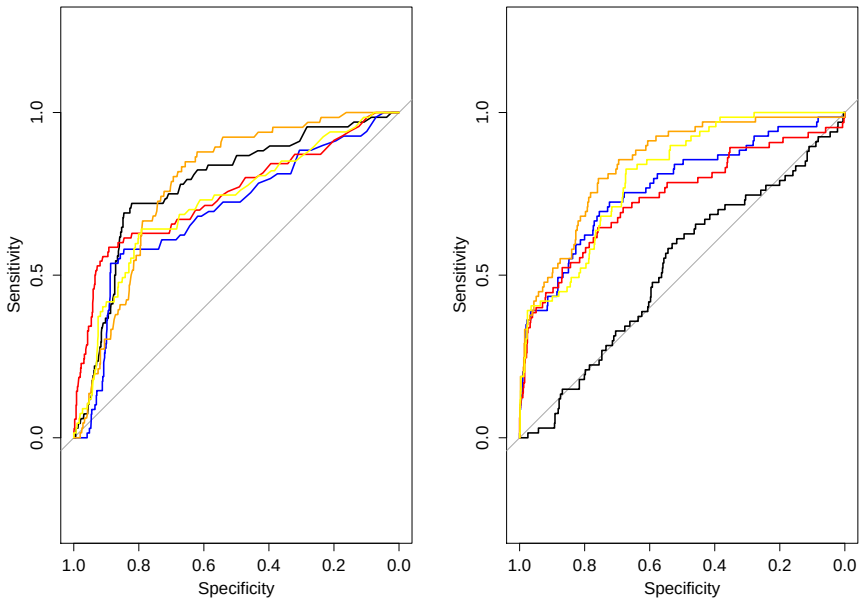
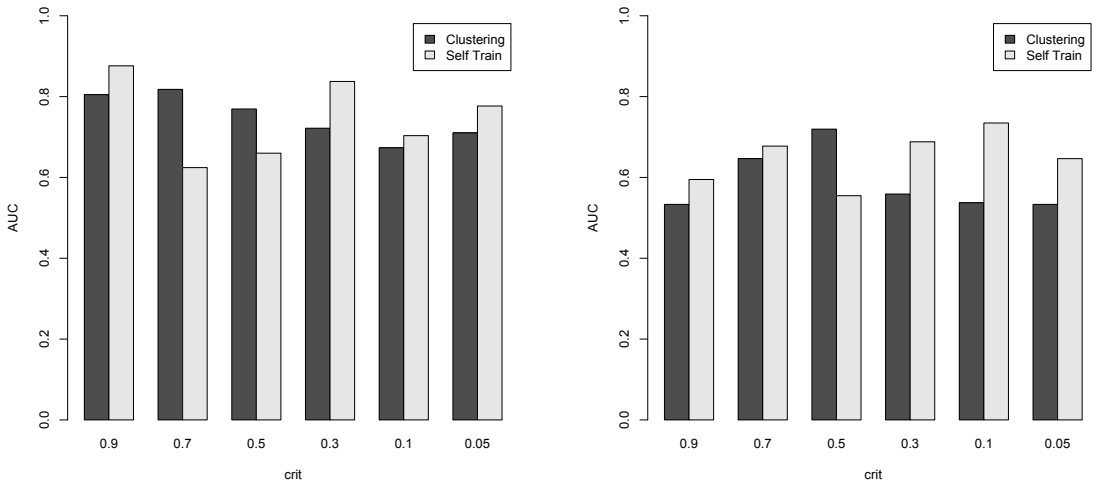


Figure 5.7: The ROC curves for the SSAP framework repeated multiple times for the drug Ibuprofen at a *crit* value of 0.1 to investigate consistency. Left Plot: Self Training, Right Plot: Clustering.

supervised techniques were applied multiple times with a *crit* value of 0.1, but the initial labels were varied. The results are displayed in Figure 5.7. It can be seen that the performance varied each time, and although the semi-supervise clustering produced good results four out of the five times, one time it performed very poorly, worse than random guessing when considering a high specificity. This is probably due to bad initial labels resulting in a poor model that then gets worse as the unlabelled data are incorporated. This is not ideal, as there is no control on the labelled data available, and applying one of the semi-supervised techniques may yield poor results for certain labelled data.

There does not appear to be an optimal value for $crit_*$ so, rather than only applying one of the algorithms, it may be optimal to apply both the semi-supervised clustering and self-trained classifier, and generate signals based on both values.



(a) Nifedipine

(b) Ciprofloxacin

(c) Ibuprofen

(d) Mean

Figure 5.8: The AUC for the ROC plots obtained by applying the semi-supervised clustering and the self training classification within the SSAP framework to the different drugs at varied *crit* values.

5.3.2.4 Summary

The results show that a semi-supervised approach that only uses data for the drug of interest generally performs well but the performance can be affected by initial errors in the self training or metric learning. The consequence is that the SSAP framework is not as consistent as the SAP framework, and although it may occasionally produce better results (depending on the initial labelled data), it may also perform very poorly. It was observed that the SAP algorithm, with the random forest classifier, had a greater AUC than either of the semi-supervised algorithms for all the *crit* values investigated for all three drugs.

Overall, the SAP algorithm produced better and more consistent results. It also has the additional benefit of not requiring knowledge of existing labels for the drug being investigated, unlike the SSAP algorithm.

5.4 Summary

In this chapter two different frameworks were proposed to signal ADRs. The frameworks were then applied to data, where the truth was known, and measures were calculated to determine the suitability of each framework. The first framework implemented a supervised algorithm and was trained using labelled data corresponding to a selection of drugs not being investigated. The second framework used a semi-supervised approach and was train using the labelled and unlabelled data for the drug of interest.

The ROC plots show that the SAP framework, using a random forest classifier, consistently generates superior results. Interestingly, using the labelled data for the drug of interest generally leads to a worse performance. Therefore, the

conclusion of this chapter is that a general model, that uses Bradford-Hill causality consideration attributes and THIN specific attributes, trained on independent drugs yields the optimal solution. The tentative results suggest such a framework is capable of signalling ADRs with a low false positive rate. It is now of interest to determine how this general model compares with existing ADR signalling algorithms and investigate if it is robust.

Chapter 6

Evaluating The ADR Learning Framework

> 'As there is no true gold standard, prospective evaluation of signal detection methods remains a challenge.'
>
> P. M. Coloma[33]

6.1 Introduction

In the previous chapters, a novel idea of automating the application of the Bradford-Hill causality considerations for mass signalling of ADRs was developed. In Chapter 4 the attributes derived from the Bradford-Hill causality considerations were presented, and used as inputs into a learning algorithm in Chapter 5. The tentative results of the novel learning algorithm, named the SAP framework, suggest that training a general classifier using knowledge of existing ADRs on attributes based on the Bradford-Hill causality considerations and THIN specific

attributes may present the opportunity to signal ADRs with a high precision and sufficiently high specificity. In this chapter, the SAP framework is evaluated by applying the specific comparison used in Chapter 3 and also by determining the frameworks ability on the HOI-DOI reference standard, as this enables a general comparison with previous and future work.

As it was hypothesised that the SAP framework will be able to generate new signals that can not be generated by existing methods, the framework will also be applied to the drug-medical event pairs for a selection of drugs that are not definitively known as ADRs or non-ADRs. The signals generated will be presented as this offers another perspective into the effectiveness of the SAP framework and may highlight new ADRs.

6.2 Motivation

The existing methods are known to suffer from the high false positive rate [156] and this means that further investigation needs to be applied to the signals that are generated. If the SAP framework has a low false positive rate, then this additional investigation will not be required, increasing the efficiency of ADR signalling. As the high false positive rate is due to signalling strong associations that are non-causal but occur due to confounding effects, the SAP framework should be more resilient to a high false positive rate as the Bradford-Hill causality considerations should help distinguish between associations due to confounding, and associations due to causation [19].

Evaluating the SAP framework on the standard reference [80] will enable other researchers to readily compare their methods with the SAP framework

and the SAP framework can be compared with previous results. However, this standard reference may be biased due to only considering a selection of HOIs. Therefore, the specific comparison is also applied to evaluate the SAP framework and determine the false positive rate when a larger number of drug-medical event pairs are analysed. If the SAP framework is shown to have a low false positive rate, but this does not inhibit its general ability, then this would be a step forward for pharmacovigilance.

6.3 Evaluation using the Standard Reference

A recent standard reference set has been introduced to enable a fair comparison between methods applied to different databases. The standard reference contains ten DOIs and nine HOIs (discussed in Chapter 2.1.5), and consists of 53 definitively known ADR or non-ADR drug-medical event pairs (9 ADRs and 44 non-ADRs). The SAP framework was applied for each of the 53 drug-medical event pairs on the THIN database and the signals generated by the framework were compared with the known truth.

Previously, the benchmark measures over all the methods and a variety of databases are an AUC of 0.77 and an AP of 0.49, the method obtaining these values had a sensitivity of 0.56, a specificity of 0.82 and a positive predictive value of 0.38 [156]. The previous comparisons have all concluded that existing methods have a high false positive rate (≥ 0.18 [156]). On the THIN database the benchmark values were a sensitivity of 0.67, a specificity of 0.68 and a precision of 0.33 [214].

6.3.1 Method

The SAP framework was evaluated by generating signals for each DOI after training the SAP framework on the other nine DOIs. As some of the DOIs have drugs in common, the drugs used during training were always excluded from the validation to prevent bias. Due to the limited number of drug-medical event pairs available for training, the random forest classifier was found to perform poorly on the reference standard, so the SAP framework with a support vector machine classifier embedded was used instead. It was also found, due to the limited training size, that feature selection was required to reduce the number of attributes used by the SAP framework. The attributes not used were the TPD filter 1 ($x_{15}^{\alpha\beta}$) and TPD filter 2 ($x_{16}^{\alpha\beta}$), LEOPARD ($x_{17}^{\alpha\beta}$), experimentation ($x_{26}^{\alpha\beta}$) and the risk different ($x_4^{\alpha\beta}$-$x_5^{\alpha\beta}$), risk ratio ($x_7^{\alpha\beta}$-$x_8^{\alpha\beta}$) and odds ratio ($x_{10}^{\alpha\beta}$-$x_{11}^{\alpha\beta}$) when only considering the first time the drug is prescribed in 13 months or the first time any drug in the same family is prescribed in 13 months.

6.3.2 Results

Table 6.1 presents the results of the signals generated using the SAP framework for the standard reference. The number of TPs was 6, the number of FPs was 7, the number of FNs (excluding the pair antibiotics and acute liver failure as that was not experienced by any patients in the subsection of the THIN database used) was 2 and the number of TNs was 37. Therefore, at its natural threshold, the SAP framework had a sensitivity of 0.75, a specificity of 0.84, a precision of 0.46 and a false positive rate of 0.16.

The general raking ability measures were a MAP (average AP) score of 0.490,

Table 6.1: The results of the SAP framework with the support vector machine classifier on the 53 standard reference set of DOIs and HOIs.

	Angi-odema	Aplastic anemia	Acute liver injury	Bleed-GI ing		Hip frac-ture	MI	Death after MI	Renal Fail-ure
ACE inhibitors	TP	TN			FP	TN			
Ampotericin	TN	TN	TN			TN		TN	TP
Antibiotics		TN	NO [1]	TN		TN	FP		TN
Antiepileptics	TN	FN			TN			TN	TN
Benzodiazapine	FP	TN	TN	FP		TP	FP		FP
Beta Blockers	TN	TN	TN		TN	TN			TN
Bisphosphonates		TN	TN		FN		TN		TN
Tricyclic Antidepressants		TN	TN	TN			TP		TN
Typical Antipsycotics					TN		TP		FP
Warfarin	TN	TN		TP		TN		TN	TN

an average AUC of 0.703 and an average P(10) of 0.2875. The DOIs antibiotics and betablockers were not used in the previous calculation due to them having no positive drug-medical event pairs, so the measures are undefined.

6.3.3 Discussion

Previous benchmarks for existing methods using the common data model on the standard reference set were an AP ranging between $0.25 - 0.49$ an AUC ranging between $0.59 - 0.77$ and a false positive rate ranging between $0.18 - 0.89$ [156]. The false positive rate of 0.16 returned by the SAP framework was lower than existing methods obtained in previous studies, but the general ranking measures were comparable with the optimal existing methods. Therefore, the results of the SAP framework using the THIN database for the standard reference set show that the

[1] This DOI-HOI pair was Not Observable (NO) using the THIN database

SAP framework is able to generate signals as well as the existing methods applied to the common data model but it has a lower false positive rate. This provides evidence to support the hypothesis that generating methods for specific database rather than the common data model may enable new signals to be generated and supports the hypothesis that incorporating knowledge of existing methods and attributes based on causation will reduce the number of false positives generated by the method.

The SAP framework was limited in this evaluation due to the small number of DOIs and HOIs resulting in a small training set. This shows that the SAP framework has even more potential, as when the training size increases, the ability of the classifier will increase and the SAP framework is likely to perform better. As the SAP framework's performance was as good, or maybe better, than existing methods when the training set was small, it is likely to significantly outperform the methods when more DOIs and HOIs are used to train the classifier. The evaluation also highlighted how adaptable the framework it, as it can use any classifier within it, so the most suitable classifier can be chosen based on the situation. Furthermore, the SAP framework only requires the classifier to be tuned and feature selection to be applied, so the number of parameters is relatively low compared to many of the existing methods, making its application more efficient.

Method	Signal Criteria	Ranking Criteria
SAP Framework	Class with most votes	Confidence of class ADR
ROR_{05}	$\text{ROR}_{05} > 1$	ROR_{05}
MUTARA_{180}	$unexlev > 0$	$unexlev$
HUNT_{180}	($unexlev$ rank)/(lev rank) > 1	($unexlev$ rank)/(lev rank)
TPD	$IC_{\Delta 05} > 0$	$IC_{\Delta 05}$

Table 6.2: The signalling and ranking criteria of the methods.

6.4 Specific Comparison

6.4.1 Method

The specific comparison, as conducted in Chapter 3.7, was repeated for the drugs Nifedipine, Ciprofloxacin, Ibuprofen, Budesonide and Naproxen and including the SAP framework as an additional method. The different criteria used by each method to generate signals or rank the pairs are described in Table 6.2. The specific comparison was chosen to be implemented in addition to evaluating the SAP on the standard reference as this enables a more rigorous evaluation of the SAP's ability to generate signals with a low false positive rate.

An additional investigating is implemented by applying the SAP framework to the 'risk' drug-medical event pairs that are not definitively known as non-ADRs or listed as ADRs on the drug packaging (i.e., the unlabelled drug-medical event pairs). This will enable potentially new ADRs to be discovered.

6.4.2 Results

6.4.2.1 Nifedipine

Natural Thresholds

Table 6.3 displays the ADR signalling methods abilities at their natural thresh-

Table 6.3: The results of the signals generated by the different ADR signalling methods applied to Nifedipine at their natural thresholds.

Method	TP	FN	FP	TN	Sensitivity	Specificity	Presicion	F-Score
SAP	22	41	8	604	0.349	0.987	0.733	0.473
ROR_{05}	44	36	164	448	0.550	0.732	0.212	0.306
$MUTARA_{180}$	54	9	267	345	0.857	0.564	0.168	0.281
$HUNT_{180}$	42	21	248	364	0.667	0.595	0.145	0.238
TPD [1]	5	58	11	601	0.079	0.982	0.313	0.127

Table 6.4: The general ranking, Area Under the Curve (AUC), partial AUC ($AUC_{[0.9,1]}$) and Average Precision (AP), results of the different ADR signalling methods applied to Nifedipine.

Method	AUC	$AUC_{[0.9,1]}$	AP
SAP	0.889	0.054	0.596
ROR_{05}	0.691	0.010	0.129
$MUTARA_{180}$	0.833	0.053	0.562
$HUNT_{180}$	0.743	0.031	0.326
TPD	0.716	0.012	0.170

old. The existing methods $MUTARA_{180}$, SRS and $HUNT_{180}$ signalled the greatest number of known ADRs, 54, 44 and 42 respectively. However, these methods also incorrectly signalled many non-ADR and had low presicion values ($0.145 - 0.212$). The SAP framework had the highest presicion, 0.733, specificity, 0.987 and F-score, 0.473. This was due to the low number of false positives.

General Ranking

Table 6.4 displays the AUC, $AUC_{[0.9,1]}$ and AP values for the five ADR signalling methods. The SAP framework had the highest AUC, $AUC_{[0.9,1]}$ and AP, with values 0.889, 0.054 and 0.596 respectively. The AUC of the SAP framework was not significantly greater than the AUC for $MUTARA_{180}$ (p-value 0.093), neither

[1] The TPD result presented was the optimal result when both the TPD_1 and TPD_2 were applied.

Table 6.5: The results of the signals generated by the different ADR signalling methods applied to Ciprofloxacin at there natural thresholds.

Method	TP	FN	FP	TN	Sensitivity	Specificity	Presicion	F-Score
SAP	13	42	0	385	0.236	1.000	1.000	0.382
ROR_{05}	19	36	71	314	0.345	0.816	0.211	0.262
$MUTARA_{180}$	55	0	327	58	1.000	0.151	0.144	0.252
$HUNT_{180}$	49	6	257	128	0.891	0.332	0.160	0.271
TPD	4	51	14	371	0.073	0.964	0.222	0.110

was the $AUC_{[0.9,1]}$ (p-value 0.471). The ROC plots are presented in Figure 6.1. The worse performing method was the ROR_{05}, with an AUC of 0.691 and an AP of 0.129.

Unlabelled Data Signals

Out 6489 unlabelled drug-medical event pairs containing Nifedipine, 233 were signalled as ADRs by the SAP framework and are displayed in Appendix D. The signals (and the number of patients experiencing them 30 days after the drug) included itching/pruritus ($\geq$1976), psoriasis (579), rash ($\geq$ 836), olecranon bursitis (483), depression ($\geq$ 3082), joint pain/arthralgia ($\geq$ 3023), appetite loss (203), tiredness (1848), excessive thirst (36), torticollis (71), dizziness (2585) and benign essential tremor (91). There were also heart related signals such as unstable angina (120) and acute myocardial infarction (114).

6.4.2.2 Ciprofloxacin

Natural Thresholds

The methods had variable sensitivities, ranging from 0.073 for $HUNT_{180}$ to 1 for $MUTARA_{180}$ and specificities, ranging from 0.0151 for $MUTARA_{180}$ to 1 for the SAP framework. This suggests their natural thresholds act at varying

Table 6.6: The general ranking, Area Under the Curve (AUC), partial AUC ($AUC_{[0.9,1]}$) and Average Precision (AP), results of the different ADR signalling methods applied to Ciprofloxacin.

Method	AUC	$AUC_{[0.9,1]}$	AP
SAP	0.812	0.048	0.614
ROR_{05}	0.713	0.007	0.190
$MUTARA_{180}$	0.851	0.042	0.547
$HUNT_{180}$	0.716	0.033	0.406
TPD	0.713	0.011	0.140

stringencies. The SAP framework was able to signal approximately 25% of the known ADRs and did not signal any non-ADRs. $MUTARA_{180}$ signalled all the 55 known ADRs but also signalled 327 non-ADRs. The TPD performed the worse for Ciprofloxacin, with the lowest F-score of 0.110 compared to the others that ranged from $0.252 - 0.382$.

General Ranking

For the drug Ciprofloxacin, $MUTARA_{180}$ had the greatest AUC, 0.851 but the SAP framework performed better when only considering a low specificity, with a $AUC_{[0.9,1]}$ of 0.048. The SAP framework also had the greatest AP value, 0.614 compared to the APs of the other methods ($0.140 - 0.547$). The AUC of $MUTARA_{180}$s ROC curve was not significantly greater than the AUC of the SAP framework ROC curve (p-values 0.241), neither was the $AUC_{[0.9,1]}$ of the ROC curve for the SAP framework compared to the $AUC_{[0.9,1]}$ of $MUTARA_{180}$s ROC curve (p-value 0.235. The ROC plots for the methods applied to signalled ADRs of Ciprofloxacin can be seen in Figure 6.1.

Unlabelled Data Signals

Table 6.7: The results of the signals generated by the different ADR signalling methods applied to Ibuprofen at there natural thresholds.

Method	TP	FN	FP	TN	Sensitivity	Specificity	Presicion	F-Score
SAP	23	52	3	783	0.307	0.996	0.885	0.455
ROR_{05}	16	59	242	544	0.213	0.692	0.062	0.096
$MUTARA_{180}$	69	6	538	248	0.92	0.316	0.114	0.202
$HUNT_{180}$	51	24	522	264	0.68	0.336	0.089	0.157
TPD	3	72	41	745	0.04	0.948	0.068	0.050

Table 6.8: The general ranking, Area Under the Curve (AUC), partial AUC ($AUC_{[0.9,1]}$) and Average Precision (AP), results of the different ADR signalling methods applied to Ibuprofen.

Method	AUC	$AUC_{[0.9,1]}$	AP
SAP	0.903	0.057	0.654
ROR_{05}	0.473	0	0.076
$MUTARA_{180}$	0.845	0.045	0.498
$HUNT_{180}$	0.595	0.020	0.196
TPD	0.654	0.002	0.102

The signals generated by the SAP framework applied to the drug Ciprofloxacin are listed in Appendix D. Out of 3574 unlabelled drug-medical event pairs containing Ciprofloxacin, 125 pairs were signalled as corresponding to ADRs. Some of the interesting signals include hypothyroidism (324), depressed mood (625), oral aphthae (285), muscle injury/strain (46), congestive heart failure (542), Incoordination symptom (807), canidial balanitis (67), confused (434), achilles tendinitis (130), left ventricular failure (318) and panic disorder (192).

6.4.2.3 Ibuprofen

Natural Thresholds

The results of the methods applied to Ibuprofen at their natural threshold are presented in Table 6.7. It can be seen that $MUTARA_{180}$ was able to signal the

majority of known ADRs (69/75) at its natural threshold but it also signalled 538 non-ADRs. The SAP framework signalled less, approximately 30% of the known ADRs (23/75), but managed to only signal 3 non-ADRs, this resulted in the SAP framework obtaining the greatest presicion, 0.885 and F-score, 0.455.

General Ranking

The SAP framework had the greatest AUC, 0.903, and $AUC_{[0.9,1]}$, 0.057, and these were significantly greater than the AUC and $AUC_{[0.9,1]}$ corresponding to the second best method $MUTARA_{180}$, with an AUC value of 0.845 and a $AUC_{[0.9,1]}$ of 0.045 (p-values 0.037 and 0.044 respectively). The SAP framework also had the greatest AP value, 0.654 compared with 0.498, 0.196, 0.102 and 0.076 corresponding to $MUTARA_{180}$, $HUNT_{180}$, TPD and the ROR_{05} respectively. These general ranking measures are contained in Table 6.8. The ROR_{05} actually performed worse than random guessing, with an AUC value under 0.5 and was unable to signal any known ADRs at a high specificity as its $AUC_{[0.9,1]}$ was 0.

Unlabelled Data Signals

When the SAP framework was applied to unlabelled data corresponding to Ibuprofen, there was a total of 200 signals out of a possible 7700. The signalled pairs included the medical events Nausea (3084), rash ($\geq$ 6155), tiredness symptom (2937), Gout (3709), essential hypertension (7883), Candidiasis (3488), Cough (1180), palpitations (1860), shortness of breath (2489), vomiting (170), patient's condition improved (22539) and myalgia (2246). A complete list of signals is contained in Appendix D.

Table 6.9: The results of the signals generated by the different ADR signalling methods applied to Budesonide at there natural thresholds.

Method	TP	FN	FP	TN	Sensitivity	Specificity	Presicion	F-Score
SAP	26	26	0	535	0.5	1	1	0.667
ROR_{05}	23	29	360	175	0.442	0.327	0.060	0.106
$MUTARA_{180}$	49	3	308	227	0.942	0.424	0.137	0.240
$HUNT_{180}$	38	14	258	277	0.731	0.518	0.128	0.218
TPD	1	51	12	523	0.019	0.978	0.077	0.031

Table 6.10: The general ranking, Area Under the Curve (AUC), partial AUC ($AUC_{[0.9,1]}$) and Average Precision (AP), results of the different ADR signalling methods applied to Budesonide.

Method	AUC	$AUC_{[0.9,1]}$	AP
SAP	0.937	0.070	0.767
ROR_{05}	0.705	0.002	0.059
$MUTARA_{180}$	0.855	0.052	0.544
$HUNT_{180}$	0.707	0.025	0.232
TPD	0.696	0.003	0.105

6.4.2.4 Budesonide

Natural Thresholds

Table 6.9 displays the results of the signals generated for Budesonide by the methods at their natural threshold. The SAP framework did not signal the most known ADRs, $MUTARA_{180}$ signalled 49 out of 52 known ADRs, but it was able to signal 50% and all the signals were correct (0 false positives). $MUTARA_{180}$ signalled 308 false positives, so only 49 out of the 357 signals generated by $MUTARA_{180}$ correspond to known ADRs. The TPD generated the least number of signals, 13 in total, and only 1 corresponded to a known ADRs, making it the wore performing method. The F-score of the SAP framework, 0.667, was over double the other methods' F-scores, in the range $[0.031, 0.240]$.

General Ranking

The SAP framework performed excellently on the drug Budesonide with an AUC of 0.937, a $AUC_{[0.9,1]}$ of 0.070 and a AP of 0.767. Both the AUC and the $AUC_{[0.9,1]}$ of the SAP framework was significantly greater than the AUC and the $AUC_{[0.9,1]}$ of $MUTARA_{180}$ (p-values 0.0126 and 0.022 respectively), the second best performing method with an AUC of 0.855 and a $AUC_{[0.9,1]}$ of 0.0524. The results for all the methods are presented in Table 6.10. The methods that obtained the lowest ranking performance were the TPD and ROR_{05}, although their AUC values were approximately 0.7, suggesting all the methods performed well for Budesonide.

Unlabelled Data Signals

There were a total of 206 signals out of a possible 5219 generated by the SAP framework when applied to unlabelled drug-medical events pairs containing Budesonide. A selection of the interesting medical events signalled as ADRs to Budesonide are micturition frequency (892), constipation (2650), pain/backache (2397), accidental falls (1513), incoordination symptom (1407), dermatitis ($\geq$ 1739), dead (125), heartburn (634), impotence (607), essential hypertension (2258), appetite loss (82), bloating (71), drug and other substances-adverse effects in therapeutic use (281), alopecia unspecified (100), tremor (201) and patient's condition worsened (927). For a list of all the signalled medical events see Appendix D.

6.4.2.5 Naproxen

Natural Thresholds

The SAP framework was able to signal approximately 40% of the known ADRs and out of the signals generated, 89% corresponded to known ADRs and only 11%

Table 6.11: The results of the signals generated by the different ADR signalling methods applied to Naproxen at there natural thresholds.

Method	TP	FN	FP	TN	Sensitivity	Specificity	presicion	F-Score
SAP	31	49	4	446	0.388	0.991	0.886	0.539
ROR_{05}	25	55	182	268	0.313	0.596	0.121	0.174
$MUTARA_{180}$	72	8	293	157	0.9	0.349	0.197	0.324
$HUNT_{180}$	54	26	274	176	0.675	0.391	0.165	0.265
TPD	6	74	15	435	0.075	0.967	0.286	0.119

Table 6.12: The general ranking, Area Under the Curve (AUC), partial AUC ($AUC_{[0.9,1]}$) and Average Precision (AP), results of the different ADR signalling methods applied to Naproxen.

Method	AUC	$AUC_{[0.9,1]}$	AP
SAP	0.883	0.055	0.700
ROR_{05}	0.510	0.000	0.136
$MUTARA_{180}$	0.793	0.036	0.503
$HUNT_{180}$	0.628	0.020	0.325
TPD	0.706	0.008	0.209

corresponded to non-ADRs. $MUTARA_{180}$ was able to signal 90% of the known ADRs, but only 20% of the total signals corresponded to ADRs, the remaining 80% were non-ADRs. The SAP framework had the greatest F-score, 0.539, with the other methods obtaining 0.324, 0.265, 0.174 and 0.119 for $MUTARA_{180}$, $HUNT_{180}$, the ROR_{05} and TPD respectively. These results are presented in Table 6.11 and the ROC plots are displayed in Figure 6.1.

General Ranking

The general ranking performance of the methods varied when applied to Naproxen, this can be seen in Table 6.12. The SAP framework and $MUTARA_{180}$ performed well, obtaining AUC values of 0.883 and 0.793 respectively and $AUC_{[0.9,1]}$ values of 0.055 and 0.036 repsectively. The SAP framework's AUC was significantly greater

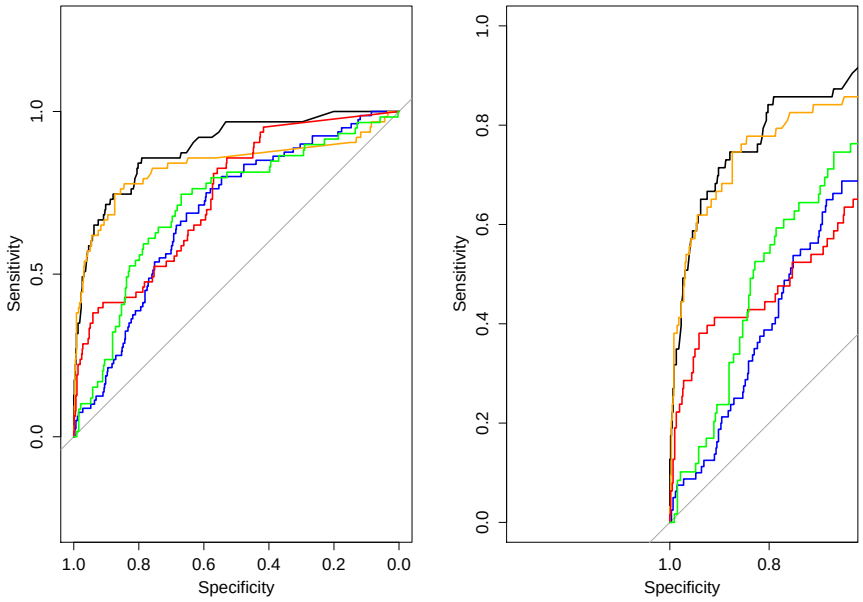
than MUTARA_{180}'s AUC (p-value 0.012), as was its $\text{AUC}_{[0.9,1]}$ (p-value 0.007). The SAP frameworks' AP was greater than the other methods, 0.7, compared with the other methods ranging from 0.136 (ROR_{05}) to 0.503 (MUTARA_{180}). The ROR_{05} performed poorly, with an AUC of 0.51, not much improvement on random guessing, and a $\text{AUC}_{[0.9,1]}$ of 0, showing it was not able to signal any known ADRs when the specificity if high.

Unlabelled Data Signals

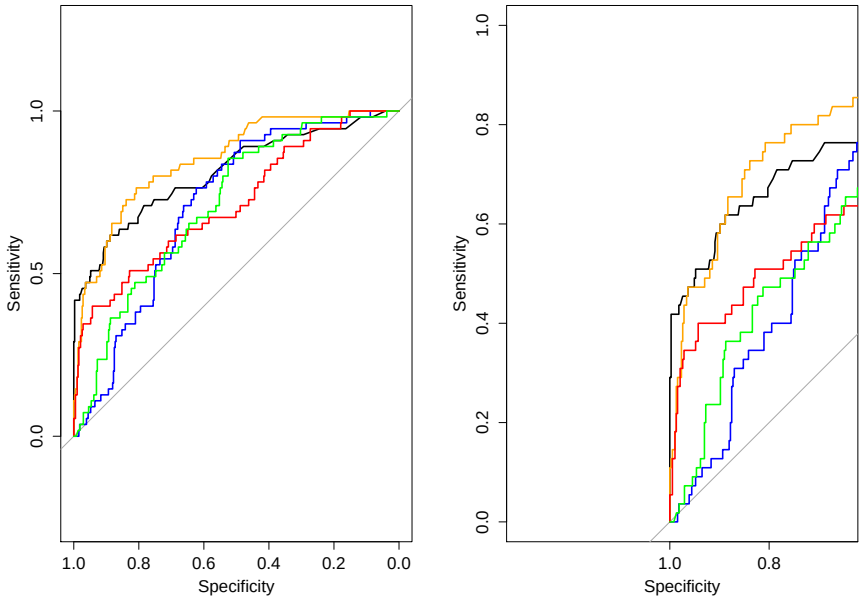
When the SAP framework was applied to the 4540 unlabelled drug-medical event pairs containing the drug Naproxen, a total of 302 pairs were signalled as corresponding to ADRs. For a list of all the medical events contained in these drug-medical event pairs see Appendix D. The medical events of interest are depression (1677), abdominal pain (1077), acquired hypothyroidism (588), anxiety states (707), breathlessness (873), hoarse (172), nausea present (232), constipation (308), unstable angina (24), vomiting (38), left ventricular failure (230), obstructive jaundice (13), acute retention of urine (13), acute non-ST segment elevation myocardial infarction (27), ocular hypertension (147), drug stopped-medical advice (529), spasms (14), congestive heart failure (368) and atria flutter (20).

6.4.3 Discussion

The SAP framework had a greater AUC value for four out of the five drugs investigated and a greater $\text{AUC}_{[0.9,1]}$ for all five drugs, compared to the existing methods. The AUC and $\text{AUC}_{[0.9,1]}$ was significantly greater, at a 5% significance

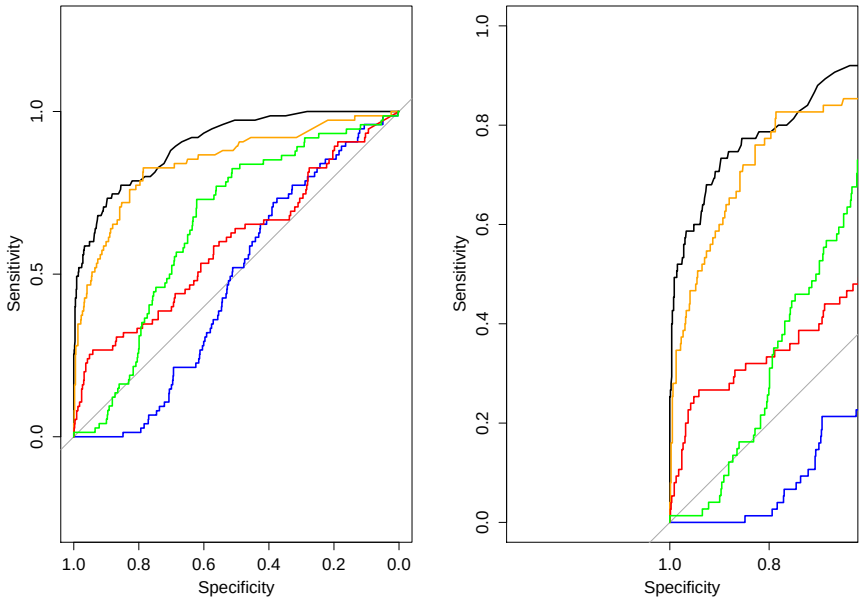


(a) Nifedipine

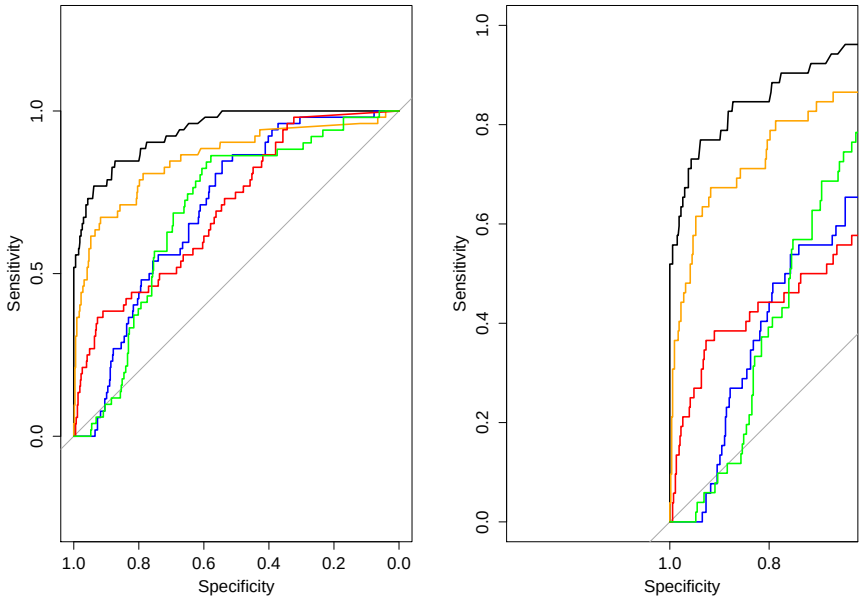


(b) Ciprofloxcin

Figure 6.1: The ROC plots for the SAP algorithm, implementing a random forest (black) and the existing methods MUTARA_{180} (orange), HUNT_{180} (red), TPD (green) and ROR_{05} (blue).

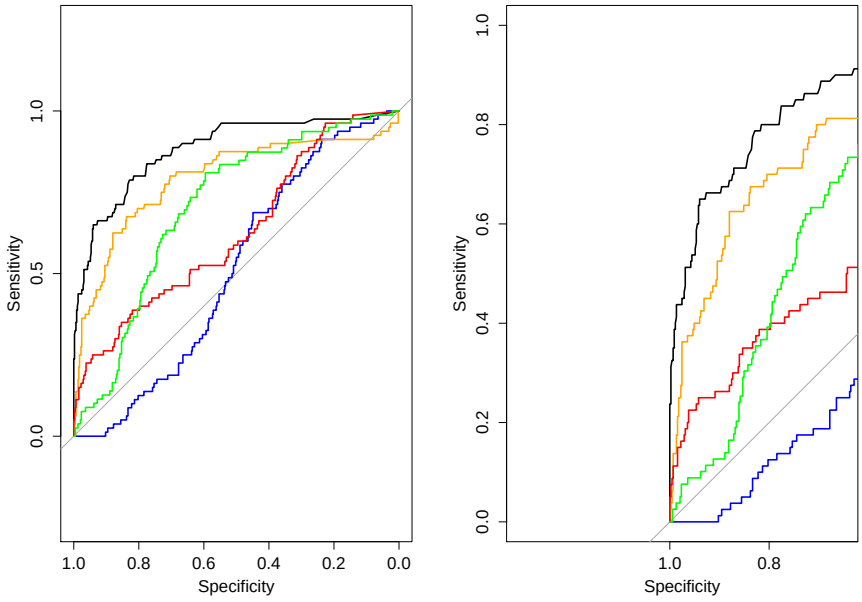


(c) Ibuprofen



(d) Budesonide

Figure 6.1: The ROC plots for the SAP algorithm, implementing a random forest (black) and the existing methods MUTARA_{180} (orange), HUNT_{180} (red), TPD (green) and ROR_{05} (blue).



(e) Naproxen

Figure 6.1: The ROC plots for the SAP algorithm, implementing a random forest (black) and the existing methods MUTARA_{180} (orange), HUNT_{180} (red), TPD (green) and ROR_{05} (blue).

level, for three of the five drugs (Ibuprofen, Budesonide and Naproxen). This suggests that the SAP framework is overall better at ranking ADRs. As the SAP framework's $AUC_{[0.9,1]}$, the AUC when the specificity is high and the false positive rate is low, was always greater than the existing methods, this shows that the SAP framework is able to signal ADRs more precisely. This is also evident by the AP score of the SAP framework being greater than the existing methods for all five drugs, resulting in an overall Mean Average Precision (MAP) score of 0.667 compared to 0.531, 0.297, 0.145 and 0.118 corresponding to $MUTARA_{180}$, $HUNT_{180}$, TPD and ROR_{05} respectively.

The SAP framework was able to signal a high percentage of ADRs while maintaining a low number of false positives. Although $MUTARA_{180}$ signalled more ADRs, it also signalled many false positives. Over all five drugs, the number of SAP and $MUTARA_{180}$ signals that were true positives were 115 and 299 respectively, but the number of false positives were 15 and 1733 respectively. Therefore, 88.5% of the SAP framework signals are likely to correspond to ADRs, while only 14.7% of $MUTARA_{180}$'s signals are likely to be ADRs. This means $MUTARA_{180}$'s natural threshold signals probably need additional filtering, whereas this is not necessary when the SAP framework is implemented.

Overall, the SAP framework managed to signal 115 out of the 325 ADRs. This corresponds to a minimum of 35.4% of the ADRs being identified, as READ code redundancy may mean that some of the 64.6% remaining non-signalled ADRs READ codes may correspond to the same medical event as the signalled READ codes. This value may be further improved by training the SAP framework on more drugs, or by adding additional attributes based on the remaining Bradford-Hill causality considerations (consistency, plausibility and coherence). The SAP

framework was also able to signal medical events that had a low frequency during the 30 day period after the drug, and medical events with a high background rate such as depression and myocardial infarction. These medical events are often difficult to signal by the existing SRS methods [70] as the association strength is often very weak.

The TPD, HUNT_{180} and ROR_{05} performed worse than the SAP framework and MUTARA_{180}. The methods were unable in general to signal ADRs without being swamped by false positives and obtained MAP scores of less than 0.5, suggesting their general ranking ability was poor on the drug-medical event pairs investigated. The TPD method may have been inhibited as it only analyses patients that have a long medical history, due to it investigating the 27 months to 21 month time period prior to the prescription. Therefore, the amount of data available for the TPD algorithm may have been smaller relative to other methods. The natural threshold of $\text{HUNT}_{180} > 1$ appeared to act at a similar stringency as MUTARA_{180} suggesting this is a good threshold to apply.

It is clear that the SAP framework was consistent over the drugs investigated and did not perform poorly on any instance. MUTARA_{180} also returned consistent results, however, the other three existing methods returned mixed results. They performed poorly for Naproxen and Ibuprofen, with the ROR_{05} being worse than random guessing and TDP performing little better.

When the SAP framework was applied to the unlabelled data corresponding to the five drugs it was able to signal many suspected ADRs and highlighted some potentially new ADRs. The results obtained from the unlabelled data were very promising but require further evaluation to confirm causality. The SAP framework successfully signalled known ADRs with obscure descriptions, and

this is additional evidence to support its ability.

6.5 Summary

The results of the SAP framework applied on the standard reference set of 53 drug-medical event pairs provide evidence that the SAP framework is able to signal ADRs with a low false positive rate. The results of applying the SAP framework on a subset of the THIN database and the results of applying existing method to the common data model are comparable. This is impressive as the common data model contains more data than the subset of the THIN database used through this research. The results provide evidence to support the argument that methods should be developed for specific databases to utilise the whole data, as it is known that information can be lost when transforming LODs into the common data model [214]. It is also clear that single databases in their raw form are useful sources for pharmacovigilance. The results also show that introducing attributes based on the Bradford-Hill causality considerations to tackle the problem of confounding can reduce the number of false positives signals.

The results of the specific comparison show that the novel SAP framework outperforms the existing methods evaluated in this thesis (MUTARA_{180}, HUNT_{180}, ROR_{05} and TPD) and signalled ADRs with a low false positive rate. The SAP framework appears to be the first ADR signalling technique that has manage to signal a sufficient number of ADRs using LODs while obtaining a low false positive rate. This is an improvement over current pharmacovigilance techniques applied to LODs and may increase the efficiency in discovering ADRs. Possible

reasons for the SAP framework's performance are the inclusion of Bradford-Hill causality consideration attributes and attributes specific to the THIN database or its ability to learn from known ADRs.

The SAP framework was able to generate new ADR signals, but further analyse needs to be performed before the the signals can be confirmed as true or not. The benefit of the SAP framework is that prior results can be used to update the framework as the signals it generates are confirmed as ADRs or non-ADRs. The SAP framework's performance should increase over time as the number of known definitive ADRs or non-ADRs increases.

Chapter 7

Conclusions

This thesis has focused on developing an ADR signalling framework, specifically for the THIN database, that can identify ADRs with a low false positive rate. It was determined that the current ADR signalling techniques, applied to the THIN database, had a high false positive rate and the majority of signals were non-ADRs. The plausible reasons for this were that the existing methods cannot distinguish between causation and association and they do not take into account the hierarchal structures embedded within the THIN database.

To overcome this issue of the methods signalling ADRs based on the strength of association rather than causation, a generalisation of the technique of considering the Bradford-Hill causality considerations to determine signals was proposed. The technique was generalised by calculating attributes based on five of the Bradford-Hill causality considerations (association strength, temporality, biological gradient, specificity and experimentation), using the THIN database, and then using knowledge of existing ADRs to find patterns embedded within the attribute values that could be used to signal ADRs. By applying a learning technique, a sixth Bradford-Hill factor, analogy, is also indirectly incorporated. Furthermore, attributes that incorporated knowledge of the THIN hierarchal structures were

also proposed and used as attributes into the learning algorithm. These attributes helped identify medical events that occurred before the drug was taken but then progressed or were recorded inconsistently.

It was shown that the SAP framework, a classifier trained using data consisting of Bradford-Hill causality considerations and THIN specific attributes corresponding to drug-medical event pairs that are known ADRs or non-ADRs, can be applied to a different drug-medical event pair to determine if the pair is an ADR or not with an specificity of 0.75 and a sensitivity of 0.84. The natural threshold false positive rate was lower than existing methods, showing that the SAP framework overcomes the current limitation of a high false positive rate that plagues the existing ADR signalling methods for LODs.

In the continuation of this chapter the contributions of this work are summarised, and suggestions are made for future directions of work to follow on from this research. The dissemination of this research is reported in the conclusion of this chapter.

7.1 Contributions

This thesis has made the following contributions:

- **Determined the benchmark for the existing methods on the THIN database**

  There is no golden standard for signalling ADRs [179] due to the lack of definitive knowledge of existing ADRs for each drug. In [214], the authors applied a selection of ADRs signalling techniques to the raw THIN database and a mapped version of the database to determine if signals are lost during

the mapping. This was the first example of the THIN database being investigated for ADR signal detection. Benchmarks for the standard reference using the raw THIN database were determined, but the authors did not apply an extensive analysis and the standard reference may contain bias.

In Chapter 3, an extensive analysis was conducted by applying a selection of existing ADR signalling methods (Reporting Odds Ratio, Temporal Pattern Discovery, Mining Unexpected Temporal Association Rules given the Antecedent and Highlighting Unexpected temporal association rules Negating Temporal association rules) to the THIN database and analysing the signals with two different perspectives. The ROR and TPD had been compared with other methods in numerous studies [156] [80] and the authors concluded that the methods performed similarly, so rather than applying all the existing methods, only these two were chosen. MUTARA and HUNT had not be incorporated in any previous comparison, so they were also added to the investigation. The previous comparisons had concluded that the methods had a high false positive rate [156] and this limited there ability.

The comparisons conducted in this research showed, consistent with previous results, that the four existing method had a high false positive rate. An interesting result was that their performance deteriorated when the number of drug-medical event pairs being investigated increased, although this may be partially due to the effect of unknown ADRs causing their results to seem worse than they are. When considering a smaller subset the drug-medical event pairs, where only definitively known ADRs or non-ADRs are

included, the benchmark AUC, $AUC_{[09,1]}$, AP, were 0.770, 0.032, 0.315 respectively. The sensitivity and specificity ranged between $0.061 - 0.894$ and $0.0366 - 0.959$ respectively.

The comparison suggested that the existing ADR signalling methods are unsuitable for signalling ADRs using the THIN database due to the large number of false positive signal generated. It would be difficult to extensively investigate each signal generated and the majority of them would be false.

- **Proposed suitable attributes to distinguish association from causation**

  The THIN database is a LOD containing prescription and medical histories for millions of patients. It offers the potential to infer temporal causal relationships between drugs and medical events, but no ADR signalling technique had been developed specifically for the THIN database. Existing methods, developed for alternative databases, determine the association strength between a drug-medical pair and signal the pairs with the greatest association. This causes a high false positive rate, as many medical events can be highly associated to a drug due to non-causal reasons. When investigating a single drug-medical event pair, researchers have often considered the Bradford-Hill causality considerations to draw conclusions [164]. As the THIN database contains data that can be used in consideration of many Bradford-Hill causality considerations, in this work, a generalisation and automatisation of this idea was proposed by extracting Bradford-Hill causality considerations based attributes from the THIN database. The attributes were then used as inputs into a learning algorithm. This is the

first attempt of such an approach.

The attributes proposed in Chapter 4 are a mixture of existing and novel calculations to cover five of the Bradford-Hill causality considerations, namely, association strength, temporality, specificity, biological gradient and experimentation. The association strength based attributes and the majority of the temporality attributes were extracted from existing pharmacovigilance methods. The specificity, biological gradient, experimentation and temporality BA ratios are all novel attributes that were developed in this work. As this work was focussing on a ADR signalling technique, specifically for the THIN database, novel attributes were also presented in Chapter 4 to deal with the hierarchal structure within the THIN data. It was concluded in Chapter 3 that the existing methods struggle with illness progressions or redundancy within the THIN database. By using the THIN medical event hierarchy, attributes were proposed that identify medical events that are more detailed or similar to medical events that were reported before the drug. Different attributes may be required for different healthcare databases, depending on any database specific issues that are identified.

- **Developed a novel supervised/semi-supervised technique for causal inference using THIN**

  After proposing the novel learning algorithm for signalling ADRs, the focus fell on what would be better, to develop a supervised classifier that is trained on labelled data corresponding to a collection drugs or to apply a semi-supervised algorithm that is trained on both labelled and unlabelled data for the drug being investigated?

In previous work, [5] and [113], researchers have trained classifiers to signal ADRs using chemical data and known ADRs. It was shown that these techniques attained a high recall and precision and the results provided evidence that incorporating knowledge of ADRs into models improves performance. The existing methods for signalling ADRs using LODs are unsupervised and do not incorporate knowledge of existing ADRs.

In Chapter 5, two learning algorithm frameworks, that use the attributes derived from the THIN data (described Chapter 4), were presented. The supervised technique, the SAP framework, applied a classifier that is trained on labelled data corresponding to various drugs. The semi-supervised technique, the SSAP framework, applied either a self-train random forest or a semi-supervised clustering technique to both the unlabelled and labelled data of a single drug. It was concluded that the SAP framework outperformed the SSAP self-train and semi-supervised approach and the SAP framework returned consistent results. This was the first attempted of implementing supervised or semi-supervised techniques to infer ADRs using a LOD.

The SAP framework consistently returned a low false positive rate, even when the training set was small. As the consuming element of the SAP framework is the training aspect, the SAP framework is highly efficient once trained and training rarely needs to occur. The SAP framework was also shown to be robust, as it was consistently able to signal ADRs with a low false positive rate over a range of drugs.

- **Evaluated the SAP framework on the THIN database**

In Chapter 6 the SAP framework was compared with the TPD, ROR, MUTARA and HUNT methods for a range of drugs using the THIN database. The results confirmed that the SAP framework, using Bradford-Hill causality considerations and THIN specific attributes and learning from known ADRs, was able to signalled ADRs with a low false positive rate, unlike the existing methods. The SAP framework obtained a greater Average Precision and $AUC_{[09,1]}$ for all the drugs investigated. The current benchmarks, set by the SAP framework, for ADR signalling methods using the THIN database are a MAP of 0.667, a sensitivity of 0.354 and a precision of 0.885. These results provided evidence to confirm the second hypothesis, that novel ADR signalling algorithms applied to the THIN database will outperform existing methods if they deal with the hierarchal structures in the THIN database, incorporate causality based attributes and learn from existing ADRs.

The SAP framework was able to generate new ADRs signals when it was applied to unlabelled drug-medical event pairs. This supports the third hypothesis. Unfortunately, additional analysis is require to confirm if the signals are true or false.

The SAP frameworks ability on the OMOP DOI-HOI standard reference containing ten DOIs and nine HOIs was limited by the training size available. However, the SAP framework's ability to generate ADR signals using the THIN database was comparable to the existing methods ability using the common data model. This is an impressive result as the common data model contains more data, and the SAP framework obtained a lower

false positive rate than existing methods. The performance of the SAP framework is likely to improve as the training size increases, so the SAP framework is likely to outcompete the existing methods when a larger standard reference set is developed. Therefore, the fourth hypothesis, that the SAP framework will outcompete the existing methods when considering the standard reference, cannot be currently confirmed but the results do provide limited evidence to support it.

7.2 Future Work

The areas of research that follow on from this research are now presented.

- **Generating attributes for the remaining Bradford-Hill causality considerations**

  In Chapter 4, attributes were developed that cover five of the nine Bradford-Hill causality considerations. The sixth, analogy, was indirectly incorporated due to using a supervised technique that looks for patterns within ADRs. The remaining considerations are consistency, plausibility and coherence. Future work could aim to generate new attributes to cover these remaining considerations. Possible suggestions for suitable attributes are, to calculate the strength of association in different databases, such as SRS databases, for the consistency factor or to incorporate attributes relating to chemical structure knowledge, such as in [113], for the coherence factor. There are two possible ideas to determine attributes for plausibility. The first idea is to mine the web, such as medical forums, and identify if the drug-medical event pair have been frequently mentioned as corresponding

to a possible ADR. In [202], the authors have used text mining techniques to identify ADRs and this idea could be adapted. The second idea is to indirectly tackle plausibility by ruling out other possibilities, this could be done by applying sequential pattern mining and filtering the explainable medical events (medical events that have progress from a prior illness).

- **Combining the SAP and SSAP frameworks using an ensemble**

  In this research a supervised framework and a semi-supervised framework were proposed in Chapter 5. Four classifiers, support vector machine, random forest, naive Bayes and logistic regression and two semi-supervised algorithms, self-trained random forest and semi-supervised k-means were applied to the data and analysed. The results showed that the random forest classifier performed the best, so this was selected and used in Chapter 6, although when the training set was small, the support vector machine classifier performed better. Future would could involved investigating an ensemble technique that uses the prediction of all the learning algorithms developed in Chapter 5 to get a final aggregated prediction.

- **Quantifying the ADRs**

  This research has produced a framework that can efficiently and precisely signal ADRs. Using this framework to signal the ADRs, the signalled ADRs could then be investigated and the additional risk of having the medical event due to taking the drug could be determined. This follow up work would add accurate quantitative information to ADRs, something that is currently lacking [171].

- **Identifying risk factors corresponding to the ADRs**

  In addition to quantifying the ADRs, the signals could also be investigated to determine risk factors. Possible methods of achieving this would be to apply association rule mining [213] to the patients' sets of medical history for all the patients taking the drug and all the patients taking the drug and experiencing the ADRs, and then identify the rules that occur more frequently in the patients experiencing the ADR.

- **Make the SAP framework run in realtime**

  The causal based attributes for each drug-medical event pair could be stored such that when new therapy and medical records are added to the database the attributes are updated. The SAP framework could then be applied to determine if the signal status of any drug-medical event pair has changed. The learning model used within the SAP framework could also be re-trained after a sufficient amount of new data is added, and could incorporate new labels as addition ADR knowledge is gained.

- **Removing the redundancy in the READ codes**

  The READ code structure has redundancy and there are multiple READ codes for the same medical event. This causes issues when trying to aggregate how frequently a medical event occurs after the drug of interest for the same population as the redundancy partitions the medical event and these partitions have smaller frequencies than if they were all grouped together. If future worked aimed to develop an algorithm that could group the READ codes that correspond to the same medical event together, the results of the ADR signalling algorithms on the THIN database would improve.

- **Adapting the framework to identify drug-drug interactions**

  Many researchers have identified the requirement of identifying drug-drug interactions ADRs. The THIN databases contains data that may be used to signal drug-drug interaction ADRs and the SAP framework can readily be adapted. Future work could aim to identify when a patient is taking two drugs within a similar time interval and then, for drug A, drug B and medical event 1, generate the attributes developed in Chapter 4 for three different prescription situations, the first would be patients only taking drug A, the second would be for patients only taking drug B and the third would be patients taking both drugs. The three sets of attributes could be combined into one data-point corresponding to drug A, drug B and the medical event 1.

7.3 Dissemination

A list of publications that have been the result of this research are listed below.

7.3.1 Journal Papers

Submitted

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**. *A Bradford-Hill causality criteria based side effect signalling framework* . Submitted to IEEE Transactions on Knowledge and Data engineering.

Accepted

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**. Signalling paediatric side effects using an ensemble of simple study designs. Drug Safety, 2014.

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**.*A novel semi-supervised algorithm for rare prescription side effect discovery.* IEEE Journal of Biomedical and Health Informatics, 2013.

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**. *Comparison of algorithms that detect drug side effects using electronic healthcare databases.* Soft Computing, 2013.

7.3.2 Conference papers

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**.*Attributes for causal inference in electronic healthcare databases.* In proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2013.

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**. *Comparing data-mining algorithms developed for longitudinal observational databases.* In Proceedings of the 12th Annual Workshop on Computational Intelligence (UKCI), 2012.

- **Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson and Richard B. Hubbard**.*Discovering sequential patterns in a UK general practice database.* In proceedings of the 1st IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2012.

- **Jenna Reps, Jan Feyereisl, Jonathan M. Garibaldi, Uwe Aickelin, Jack E. Gibson, Richard B. Hubbard**. *Comparing data-mining algorithms developed for longitudinal observational databases.* In Proceedings of the 11th Annual Workshop on Computational Intelligence (UKCI), 2011.

- **Feng Gu, Jan Feyereisl, Robert Oates, Jenna Reps, Julie Greensmith, Uwe Aickelin**. *Quiet in Class: Classification, Noise and the Dendritic Cell Algorithm.* In Proceedings of the 10th International Conference on Artificial Immune Systems (ICARIS 2011), 2011.

Appendix A

The THIN Database

Introduction

The THIN database is a longitudinal resource containing temporal medical data corresponding to over 3.5 million active patients and 11.5 million total patients. The data are anonymously extracted from each individual general practice's Vision clinical system, validated and combined to generate the THIN database. The current database is 326Gb and covered 6.05% of the UK in 2012, with over 0.6 billion medical records (i.e., entries detailing an instance of a medical event such as an illness, observation or laboratory event) and approximately 1 billion therapy records (i.e., entries detailing an instance of a drug prescription). There is a slightly higher relative proportion of female patients than male patients in the database, with 47.7% of a patients being male and 52.3% being female, whereas the 2011 census suggests the UK population is 49.1% male and 50.9% female. The number of general practices included within the database is expanding over time, with 12 new practices recruited during the first three quarters of 2013.

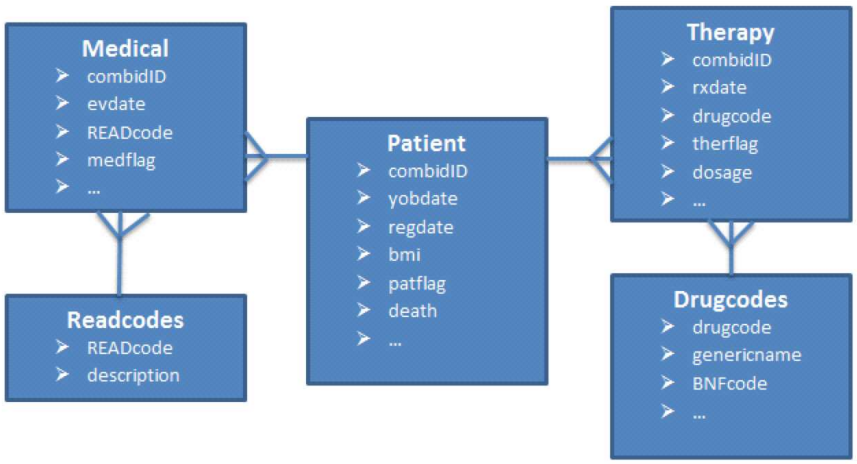


Figure A.1: An entity relationship diagram of the THIN database.

The database is also expanding due to recently occurring records from registered practices being added over time.

Structure

The structure of the main THIN database is illustrated in Figure A.1, there are additional tables not included into the diagram due to them not being incorporated within this research. The three main tables are the patient table, the therapy table and the medical table, see Figures A.2-A.4. Each patient within the THIN database is represented by a unique anonymous patient id, named the combid, and the patient table contains the attributes of each patient (e.g., their year of birth, their body mass index, their smoking habits, the year they registered and the date of death if they have died). The medical table stores the temporal data regarding the patients' medical events. Each entry in the medical table contains a combid that refers to the patient experiencing the medical event,

a READ code that corresponds to a medical event and the date that the medical event occurred. The READ codes are there due to database normalisation, but one advantage of using the READ codes rather than string descriptions to represent a medical event is that they have a hierarchical structure that may be useful when applying data analysis. The READ codes and their structure are discussed in greater detail further in thin Chapter. The therapy table contains records regarding drug prescriptions. Each therapy record contains the combid referring to the patient being prescribed the drug, a drugcode corresponding to the drug being prescribed and the prescription date. The drugcode is also introduced due to database normalisation. The drugcode does not have an obvious structure but each drugcode is linked to up to three British National Formulary (BNF) codes corresponding to the main chemical components that make up the drug. The BNF codes do have a hierarchical structure and can be used to identify similar groups of drugs. The BNF codes are also discussed in greater detail in the latter section of this Chapter.

READ Codes

The READ codes are a clinical terminology thesaurus used for recording medical events within General Practice databases. Each medical event is encoded into a READ code, and the READ code consists of five elements from the alphabet $\{1-9, a-z, A-Z, \bullet\}$. The READ codes have a hierarchal tree structure with five levels. The medical events become more specific as the level increases, so the child READ codes correspond to the same medical event as their parent READ code but are more specific. The level of a READ code $\mathbf{x} = x_1x_2x_3x_4x_5$ is calculated

Results | Messages

	combid	prac	patid	patflag	yobstring	hh	sex	regdate	regreal	xferdate	xferreal	regrea	death
1	h998101AD	h9981	01AD	A	19830000	001455	1	19880921	1988-09-21 00:00:00.000	19890727	1989-07-27 00:00:00.000	03	000(
2	h998101aD	h9981	01aD	A	19420000	003428	1	20001227	2000-12-27 00:00:00.000	00000000	NULL	00	000(
3	h998101ad	h9981	01ad	A	19710000	001646	1	19901017	1990-10-17 00:00:00.000	19940518	1994-05-18 00:00:00.000	02	000(
4	h998101ae	h9981	01ae	A	19470000	001646	1	19901017	1990-10-17 00:00:00.000	20020123	2002-01-23 00:00:00.000	03	000(
5	h998101af	h9981	01af	A	19830000	001646	1	19901017	1990-10-17 00:00:00.000	20020123	2002-01-23 00:00:00.000	03	000(
6	h998101aG	h9981	01aG	A	19140000	001189	2	19950501	1995-05-01 00:00:00.000	20001010	2000-10-10 00:00:00.000	03	000(
7	h998101ag	h9981	01ag	A	19730000	002868	1	19901017	1990-10-17 00:00:00.000	00000000	NULL	00	000(
8	h998101Ah	h9981	01Ah	A	19480000	000717	1	19950222	1995-02-22 00:00:00.000	20040521	2004-05-21 00:00:00.000	01	200(
9	h998101ah	h9981	01ah	A	19480000	001646	2	19901017	1990-10-17 00:00:00.000	20020123	2002-01-23 00:00:00.000	03	000(
10	h998101ai	h9981	01ai	A	19360000	001360	1	19880111	1988-01-11 00:00:00.000	19950203	1995-02-03 00:00:00.000	03	000(
11	h998101Aj	h9981	01Aj	A	19490000	003003	2	19980710	1998-07-10 00:00:00.000	20041105	2004-11-05 00:00:00.000	02	000(
12	h998101aj	h9981	01aj	A	19140000	001267	1	19880113	1988-01-13 00:00:00.000	19890210	1989-02-10 00:00:00.000	01	198(
13	h998101ak	h9981	01ak	A	19170000	001267	2	19871221	1987-12-21 00:00:00.000	20070717	2007-07-17 00:00:00.000	27	000(
14	h998101Al	h9981	01Al	A	19000000	000076	1	19820825	1982-08-25 00:00:00.000	19920331	1992-03-31 00:00:00.000	02	000(
15	h998101Am	h9981	01Am	A	19150000	000927	1	19820705	1982-07-05 00:00:00.000	19900331	1990-03-31 00:00:00.000	01	198(
16	h998101aM	h9981	01aM	A	19170000	001442	2	19880818	1988-08-18 00:00:00.000	19940124	1994-01-24 00:00:00.000	02	000(
17	h998101am	h9981	01am	A	19360000	002869	2	19971009	1997-10-09 00:00:00.000	20040112	2004-01-12 00:00:00.000	27	000(
18	h998101AO	h9981	01AO	A	19280000	000899	2	19590511	1959-05-11 00:00:00.000	20060206	2006-02-06 00:00:00.000	03	000(
19	h998101AP	h9981	01AP	A	19960900	002994	1	19980619	1998-06-19 00:00:00.000	19981202	1998-12-02 00:00:00.000	02	000(
20	h998101ap	h9981	01ap	A	19420000	003457	1	20010323	2001-03-23 00:00:00.000	20091104	2009-11-04 00:00:00.000	27	000(
21	h998101aS	h9981	01aS	A	19690000	000051	2	19830708	1983-07-08 00:00:00.000	19881212	1988-12-12 00:00:00.000	03	000(

Figure A.2: A screen shot of the patient table contained within the THIN database.

Results | Messages

	combid	prac	patid	rxdate	rxdatereal	drugcode	therflag	doscode	rxqty	rxdays	private	staffid	rxtype	opno
1	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	93619997	Y	0000472	56.00000	000	N	0008	1	00000.00
2	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	86989998	Y	0000200	56.00000	000	N	0008	1	00000.00
3	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	96277997	Y	0012382	1.000000	000	N	0009	1	00000.00
4	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	98776998	Y	0000929	112.0000	000	N	0009	1	00000.00
5	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	96940997	Y	0000200	56.00000	000	N	0003	1	00000.00
6	a6732010h	a6732	010h	19990707	1999-07-07 00:00:00.000	93619996	Y	0000472	56.00000	000	N	0008	1	00000.00
7	a6732010h	a6732	010h	19990729	1999-07-29 00:00:00.000	98815990	Y	0000001	1.000000	000	N	0009	1	00000.00
8	a6732010h	a6732	010h	19990729	1999-07-29 00:00:00.000	96940997	Y	0000200	56.00000	000	N	0003	1	00000.00
9	a6732010h	a6732	010h	19990811	1999-08-11 00:00:00.000	96277997	Y	0012382	1.000000	000	N	0009	1	00000.00
10	a6732010h	a6732	010h	19990824	1999-08-24 00:00:00.000	86990998	Y	0000200	56.00000	000	N	0008	1	00000.00
11	a6732010h	a6732	010h	19990824	1999-08-24 00:00:00.000	93619997	Y	0000472	56.00000	000	N	0008	1	00000.00
12	a6732010h	a6732	010h	19990824	1999-08-24 00:00:00.000	96940997	Y	0000200	56.00000	000	N	0003	1	00000.00
13	a6732010h	a6732	010h	19991027	1999-10-27 00:00:00.000	96329998	Y	0000447	112.0000	000	N	0009	1	00000.00
14	a6732010h	a6732	010h	19991112	1999-11-12 00:00:00.000	93619997	Y	0000472	56.00000	000	N	0008	1	00000.00
15	a6732010h	a6732	010h	19991112	1999-11-12 00:00:00.000	98776998	Y	0000929	112.0000	000	N	0009	1	00000.00
16	a6732010h	a6732	010h	19991112	1999-11-12 00:00:00.000	96940997	Y	0000200	56.00000	000	N	0003	1	00000.00
17	a6732010h	a6732	010h	19991210	1999-12-10 00:00:00.000	89385998	Y	0000200	28.00000	000	N	0008	1	00000.00
18	a6732010h	a6732	010h	19991210	1999-12-10 00:00:00.000	86990998	Y	0000200	56.00000	000	N	0008	1	00000.00
19	a6732010h	a6732	010h	19991210	1999-12-10 00:00:00.000	86990998	Y	0000200	56.00000	000	N	0008	1	00000.00
20	a6732010h	a6732	010h	20000124	2000-01-24 00:00:00.000	96940997	Y	0000200	56.00000	000	N	0003	1	00000.00
21	a6732010h	a6732	010h	20000124	2000-01-24 00:00:00.000	86990998	Y	0000200	56.00000	000	N	0008	1	00000.00

Figure A.3: A screen shot of the therapy table contained within the THIN database.

Results | Messages

	combid	prac	patid	evdate	evdatereal	enddate	enddatereal	dtype	medcode	medflag	staffid	source	episode	nhsspec
1	[illegible]	[illegible]	00??	20061227	2006-12-27 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	0004	0	0	000
2	[illegible]	[illegible]	00??	20061228	2006-12-28 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000C	0	0	000
3	a67060??	a6706	00??	20061228	2006-12-28 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000C	0	0	000
4	a67060??	a6706	00??	20061228	2006-12-28 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000C	0	0	000
5	a67060??	a6706	00??	20061228	2006-12-28 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000C	0	0	000
6	a67060??	a6706	00??	20080725	2008-07-25 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000b	0	0	000
7	a67060??	a6706	00??	20080725	2008-07-25 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000b	0	0	000
8	a67060??	a6706	00??	20080725	2008-07-25 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000b	0	0	000
9	a67060??	a6706	00??	20080725	2008-07-25 00:00:00.000	00000000	NULL	01	ZZZZZ00	R	000b	0	0	000
10	a67060??	a6706	00??	20080901	2008-09-01 00:00:00.000	00000000	NULL	01	9N36.00	R	000L	0	4	000
11	a67060??	a6706	00??	20080915	2008-09-15 00:00:00.000	00000000	NULL	01	9N36.00	R	000L	0	4	000
12	a67060??	a6706	00??	20080915	2008-09-15 00:00:00.000	00000000	NULL	11	G65..00	R	0004	0	0	000
13	a67060??	a6706	00??	20080915	2008-09-15 00:00:00.000	00000000	NULL	01	G65..00	R	0004	0	4	000
14	a67060??	a6706	00??	20080923	2008-09-23 00:00:00.000	00000000	NULL	01	66X..00	R	0004	0	4	000
15	a67060??	a6706	00??	20080923	2008-09-23 00:00:00.000	00000000	NULL	01	9N36.00	R	000L	0	4	000
16	a67060??	a6706	00??	20080926	2008-09-26 00:00:00.000	00000000	NULL	01	9N2S.00	R	0002	0	4	000
17	a67060??	a6706	00??	20081020	2008-10-20 00:00:00.000	00000000	NULL	01	9N36.00	R	000L	0	4	000
18	a67060??	a6706	00??	20081223	2008-12-23 00:00:00.000	00000000	NULL	01	9N36.00	R	000L	0	4	000
19	a67060??	a6706	00??	20090228	2009-02-28 00:00:00.000	00000000	NULL	01	9NL..00	R	000H	0	4	000
20	a67060??	a6706	00??	20090731	2009-07-31 00:00:00.000	00000000	NULL	01	1M10.00	R	0004	0	4	000
21	a67060??	a6706	00??	20090827	2009-08-27 00:00:00.000	00000000	NULL	01	9N0M.00	R	000D	0	4	000

Figure A.4: A screen shot of the medical table contained within the THIN database.

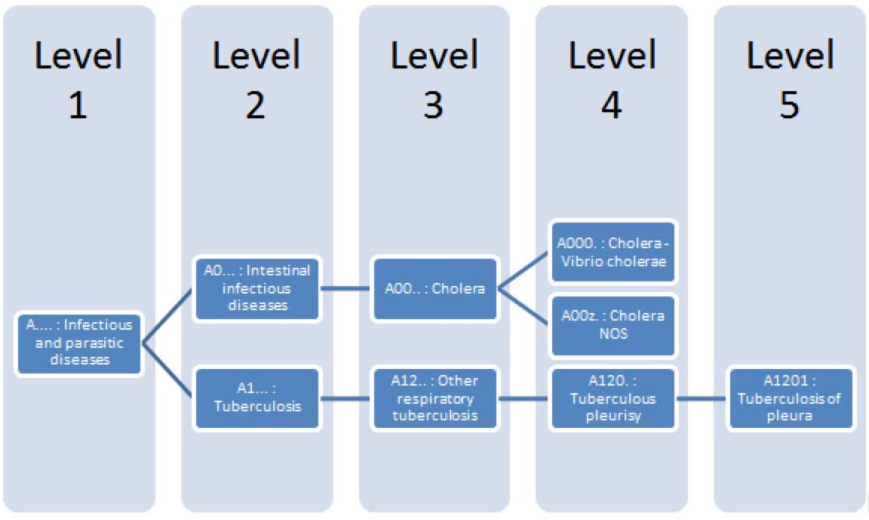


Figure A.5: An example of the branch of the THIN READ code tree.

as,

$$Lv(\mathbf{x}) = \begin{cases} \arg\min_i\{(i-1)|x_i = \bullet\} & \text{if } \exists i \text{ s.t. } x_i = \bullet \\ 5 & \text{otherwise.} \end{cases} \tag{A.1}$$

An example of a branch in the READ codes is,

A••••• Infectious and parasitic diseases (level 1)

A1••• Tuberculosis (level 2)

A12•• Other respiratory tuberculosis (level 3)

A120• Tuberculosis pleurisy (level 4)

A1201 Tuberculosis of pleura (level 5)

where it can be seen that all the READ codes above are infections and the infection represented by the READ code becomes more detailed as the level increases. A graphical illustration of this section of the READ code tree can be seen in Figure A.5.

Unfortunately, the READ codes have redundancies and a single medical event may have multiple corresponding READ codes found in widely varying branches of the READ code tree. This can lead to issues during data analysis as it is difficult to aggregate the data for the READ codes corresponding to the same medical event, and the partitioning can result in a lower confidence in the results that are obtained. There are also problems with inconsistent READ code usage by medical staff. For example, some staff may frequently enter high level specific READ codes while others may have a tendency to enter low level READ codes that are less specific. Furthermore, it is common to find 'temporal READ code progressions', where a low level READ code is initially recorded and shortly in

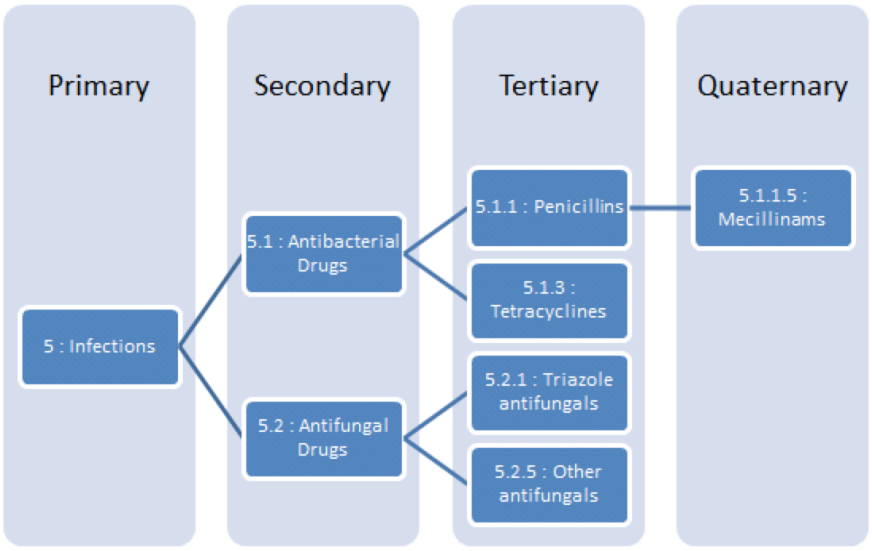


Figure A.6: An example of the branch of the British National Formulary (BNF) tree.

time afterwards a child or grand-child READ code is recorded due to additional knowledge being obtained.

BNF Codes

The BNF codes are based on BNF sections. They have a hierarchal tree structure linking drugs that are prescribed for the similar indication (i.e., the reason for being given the drug), and drugs with the same BNF code are from the same drug family. Figure A.6 illustrates a branch of the BNF code tree. If we consider each BNF code to be represented by $\mathbf{y}_i = y_{i1}.y_{i2}.y_{i3}.y_{i4}$, where each element is in the alphabet $\{1-15, 00\}$, then y_{i1} is the primary category, y_{i2} is the secondary category, y_{i3} is the tertiary category and y_{i4} is the quaternary category. There are 15 different primary categories, these primary categories relate to the the most general description of the drug indication. The similarity between two BNF codes

y_{i1}	Category
1	Gastro-intestinal system
2	Cardiovascualar system
3	Respiratory system
4	Central nervous system
5	Infections
6	Endocrine system
7	Obstetrics, gynaecology, and urinary-tract disorders
8	Malignant disease and immunosuppression
9	Nutrition and blood
10	Musculoskeletal and joint diseases
11	Eye
12	Ear, nose, and oropharynx
13	Skin
14	Immunological products and vaccines
15	Anaesthesia

can be calculated as,

$$S(\mathbf{y}_i, \mathbf{y}_j) = \frac{|\{y_{ik}|y_{ik} = y_{jk}\}|}{max(|\{k|y_{ik} \neq 00\}|, |\{k|y_{jk} \neq 00\}|)} \tag{A.2}$$

where the similarity measure is 1 if and only if the two BNF codes are the same, and greater than zero if and only if the BNF codes correspond to drugs prescribed for a similar indication.

Issues & Validation

There are known issues with the database including concept drift and problems with the level of time stamp detail. In general, the data is validated during extraction and additional fields are added into the tables to indicate the integrety of each record, so problematic records can be excluded from the study.

Data Collection Issues

One of the main limitations of the THIN data is changes in the way data is collected or the type of data collected over time may lead to concept drift. Over time the READ codes that are actively used may change, new READ codes may get added and old READ codes may be removed. For example, it is common for old records to contain the READ code 'ZZZZZ' corresponding to an unmappable medical event. The drug prescription rate is unlikely to stay constant over time, as new knoweldge of suspected ADRs or new studies detailing the effectiveness of a drug can impact a General Practioneers decision to prescribed a drug. It is also common for new drugs to be introduced.

Time Stamps

Each record in the medical and therapy table contains a time stamp. These time stamps are the day that the doctors entered the event of prescription into the database. As the time stamp is only in days, it is not possible to determine the order for the medical events and prescriptions within one day. When a medical event and prescription are recordered for the same paient on the same day it may be possible that the patient was prescribed the medication due to the medical event or that the medical event is an adverse drug reaction of the medication.

To address the uncertainty of the order of events with the same timestamp for the same patient, the medical events recorded on the day a drug is prescribed are often ignorred from the calculation of assocation between a drug and medical event.

Appendix B

Drugs

Drugs Investigated

NSAIDs

The drugs Ibuprofen, Ketoprofen, Fenoprofen and Celecoxib used in this study are all from the same drug family known as non-steroidal anti-inflammatory drugs (NSAIDs). These drugs are typically prescribed for continuous pain associated with inflammation and have a variety of common side effects including gastrointestinal disturbances, hypersensitivity reactions and depression. Rarer side effects include congestive heart failure, renal failure and hepatic failure. Elderly patients are more prone to side effects associated with NSAIDs. In this study the the drugs tended to be prescribed sightly more to females with the male proportion ranging from $0.335 - 0.405$ and to older patients, although Ibuprofen was prescribed to younger patients more than the other NSAID drugs. The NSAID drug prescribed the most was Ibuprofen with over a million first in 13 month

Table B.1: Information about the NSAID drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
celecoxib	68036	62946	62100	63416	62.49	0.335
ibuprofen	1178163	1012555	858819	903415	45.56	0.405
ketoprofen	72946	65718	61710	63536	58.17	0.375
fenoprofen	1255	1008	975	1036	56.29	0.404

prescriptions, whereas Fenoprofen was only prescribed 1225 times for the first time in 13 months, see Table B.1.

Quinolones

The quinolones are a class of drugs used to treat bacterial infections such as respiratory track infections and urinary-track infections. Ciprofloxacin, levofloxacin, moxifloxacin, nalidixic acid and norfloxacin are drugs from the quinolone family that are investigated in this paper. The quinolones have many side effects, including tendon rupture. The average age of the patients prescribed the quinolones for the first time in 13 months was similar between all the drugs, around the late fifties. The male proportion shows that females are prescribed quinolones more than males, but this was more obvious for norfloxacin and nalidixic acid. Ciprofloxacin was the most prescribed quinolone and moxifloxacin was the least common, with only 1465 prescriptions. Table B.2 shows the information on the drugs from the THIN database.

Table B.2: Information about the Quinolone drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
ciprofloxacin	280011	250158	227739	235420	55.64	0.440
levofloxacin	7662	7028	6775	6928	60.55	0.43
norfloxacin	14876	13224	12220	12625	56.83	0.262
moxifloxacin	1465	1347	1343	1371	62.09	0.419
nalidixic acid	4273	3646	3620	3787	55.63	0.127

Table B.3: Information about the tricyclic drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
doxepin	6752	6029	5908	6104	56.69	0.316
lofepramine	45532	38565	37642	39517	51.39	0.285
nortriptyline	11775	10519	10307	10650	54.43	0.286

Tricyclic Antidepressants

Tricyclic antidepressant drugs are a family of drugs used to treat depression and are known to cause, among others, cardiovascular and central nervous system side effects. The three drugs, doxepin, lofepramine and nortriptyline where selected in this paper. The tricyclic antidepressants investigated are prescribed to patients with similar ages and genders and tend to be prescribed more often to older females. The main difference between the drugs is that doxepin is only prescribed to 6752 patients whereas the other two drugs are prescribed to more than 10000 patients, see Table B.3.

Table B.4: Information about the calcium channel blocker drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
nifedipine	125491	112715	112499	115823	65.29	0.453
verapamil	24334	22000	21896	22513	65.01	0.405
felodipine	69534	65093	64036	65202	67.46	0.454
amlodipine	270918	251316	249972	254876	66.68	0.494
nicardipine	2796	2510	2511	2593	65.91	0.481

Calcium Channel Blockers

The drugs nifedipine, nicardipine, amlodipine, felodipine and verapamil are all calcium channel blocker that are used to treat high blood pressure and raynaud's phenomenon. It is common for the calcium channel blockers to be prescribed with other drugs and applying the existing algorithms to detect side effects on the calcium channel blockers will investigate the effect of confounding due to multiple prescriptions. The drug nifidipine was previously used to investigate the TPD applied to the UK IMA Disease Analyzer, so investigating the calcium channel blockers will also give insight into how robust the TPD is when applied to different electronic healthcare databases. The calcium channel blockers are generally prescribed for the first time in 13 months to patients around 65 years old. Amlodipine and nicardipine are prescribed only slightly more to females than males, whereas the other calcium channel blockers investigated are prescribed even more often to females. Amlodipine and nifedipine have been prescribed over 100000 times for the first time in 13 months in the THIN database, but nicardipine has only been prescribed 2796 times, see Table B.4.

Table B.5: Information about the sulphonylurea drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
glibenclamide	11874	10356	10377	10768	65.12	0.540
gliclazide	45824	41626	40537	41612	65.02	0.546
glimepiride	10957	10156	9882	10081	64.20	0.534
glipizide	5315	4856	4614	4731	66.50	0.535
tolbutamide	3113	2758	2793	2894	69.40	0.487

Sulphonylureas

The sulphonylurea drug family includes tolbutamide, glibenclamide, gliclazide, glimepiride and glipizide. They are a class of antidiabetic drugs used for the management of type 2 diabetes mellitus. The sulphonylureas are prescribed for the first time in 13 months to older patients will an average age around 65 years old and all the sulphoylureas investigated except tolbutamide are prescribed more often to males, with approximately equal male proportions. Glipizide and tolbutamide are the less frequently prescribed sulphonylurea drugs. The general information about each of the sulphonylurea drugs can be seen in Table B.5.

Penicillins

The last drug family is the Penicillin drugs amoxicillin, ampicillin, flucloxacillin, benzylpenicillin and phenoxymethlypenicillin. These drugs are used to treat bacterial infections. The number of times the drugs are recorded as being prescribed in the THIN database varies between 2000 to over two million. There is also a divergence between the average age of the patients prescribed each of the drugs,

Table B.6: Information about the penicillin drugs investigated in this paper. Total is the number of times the drug is prescribed for the first time in 13 months, age is the average age of the patients who are prescribed the drug for the first time in 13 months and male proportion is the number of patients that are male divided by the total number of patients who are prescribed the drug for the first time in 13 months.

Drug	Total	TPD	MUTARA	ROR	Age	male proportion
amoxicillin	2795759	2321098	1593874	1718875	38.84	0.427
benzylpenicillin	2071	1610	1840	1972	31.79	0.471
flucloxacillin	971174	834017	729967	765428	41.42	0.456
phenoxymethly	55397	45941	45679	48142	29.67	0.396
ampicillin	80655	63458	64827	69381	39.18	0.423

with the penicillins generally being prescribed to younger patients than many of the other drugs families investigated in this paper. The male proportion is fairly similar between the different penicillin drugs, with females being prescribed the drug more often than males, see Table B.6.

Appendix C

Software Details and Preliminary Work

C.1 Software Details

The data were stored in MS SQL server and the data manipulation (generation of the Bradford Hill causality consideration attributes) was performed using SQL. The classification was performed using the function 'train' and the feature selection used to pre-process the data prior to classification for all the classifiers expect random forest was the function 'rfe' within the 'caret' package [97] in the open source software R. The 'rfe' function found the subset of attributes that maximised the accuracy of the classification. The 'train' function trained the various classifiers based on maximising the AUC performance measure using a parameter grid search.

C.2 Wrapper Feature Selection

Table C.1: The features selected and their rank of importance based on applying naive Bayes wrapper for the analysis performed in Chapter 5.

Attribute	Nifedipine	Ciprofloxacin	Ibuprofen
Subset Size	25	30	30
TPD IC delta (x_1)	✔(17)	✔(14)	✔(7)
TPD IC delta 95% CI (x_2)	✔(22)	✔(2)	✔(18)
RD all (x_3)	×	✔(1)	✔(3)
RD first drug (x_4)	✔(19)	✔(3)	✔(1)
RD first BNF (x_5)	×	✔(6)	✔(2)
RR all (x_6)	✔(21)	✔(29)	✔(10)
RR first drug (x_7)	✔(10)	✔(27)	✔(5)
RR first BNF (x_8)	✔(5)	✔(24)	✔(12)
OR all (x_9)	✔(25)	✔(30)	✔(9)
OR first drug (x_{10})	✔(9)	✔(28)	✔(6)
OR first BNF (x_{11})	✔(4)	✔(25)	✔(11)
AB month all (x_{12})	✔(16)	✔(23)	✔(19)
AB month first drug (x_{13})	✔(7)	✔(16)	✔(25)
AB month first BNF (x_{14})	✔(3)	✔(20)	✔(27)
TPD filter 1 (x_{15})	✔(15)	×	✔(29)
TPD filter 2 (x_{16})	×	×	×
LEOPARD (x_{17})	×	✔(18)	✔(26)
Read code Lv 5 (x_{18})	✔(18)	✔(21)	✔(28)
Age all (x_{19})	✔(8)	✔(7)	✔(15)
Age first drug (x_{20})	✔(2)	✔(5)	✔(8)
Age first BNF (x_{21})	✔(1)	✔(4)	✔(4)
Gender all (x_{22})	×	✔(11)	×
Gender first drug (x_{23})	✔(11)	✔(12)	✔(17)
Gender first BNF (x_{24})	✔(6)	✔(13)	✔(21)
Dosage (x_{25})	✔(13)	✔(26)	✔(20)
Experimentation (x_{26})	×	✔(17)	✔(13)
Noise (x_{27})	✔(24)	✔(19)	✔(14)
Illness progression (x_{28})	×	✔(15)	✔(24)
AB month Lv 3 (x_{29})	✔(23)	✔(9)	✔(16)
AB month Lv 4 (x_{30})	✔(14)	✔(22)	✔(23)
Read code Lv 4 (x_{31})	✔(12)	✔(10)	×
Read code Lv 3 (x_{32})	✔(20)	✔(8)	✔(22)
Read code Lv 2 (x_{33})	×	×	✔(30)

C.3 Preliminary Work

The following is extracted from my conference paper title 'Attributes for causal inference in longitudinal observational databases':

Feature Selection

In this study we apply a multivariate filter, the Correlation-based Feature Selection (CFS) algorithm [69], as this algorithm is not dependent on a specific classifier. The CFS algorithm finds the optimal feature subset based on the trade-off between how correlated the class labels are to the feature subset and how intercorrelated the features of the subset are.

The feature selection was applied to the attributes descirbed in Tables C.2-C.3. The data used in this study are extracted from The Health Improvement Network database (www.thin-uk.com) and can be found at: http://www.ima.ac.uk/reps.

Results

Table C.4 shows that the optimal attribute subset to use for ADR discovery is LEOPARD, RD_{13BNF}, ABratio Level 3, Gender Ratio and Read Code Level. The temporal and strength attributes had the greatest correlation with the class labels, whereas 75% of the dosage attributes has a zero correlation measure.

Discussion

The results show that the temporal and strength attributes are key for signalling ADRs as these had the highest correlation with the class labels but the specificity attributes Gender Ratio and Read Code level offered potentially new in sight than

Table C.2: Attribute Summary Table

Feature	Criterion	Description
RR, RD, OR	Strength	The Risk Ratio, Risk Difference and Odds Ratio for all prescriptions.
RR_{13d},RD_{13d},OR_{13d}	Strength	The Risk Ratio, Risk Difference and Odds Ratio for drugs prescribed for the first time in 13 months.
RR_{13BNF},RD_{13BNF} ,OR_{13BNF}	Strength	The Risk Ratio, Risk Difference and Odds Ratio for drugs corresponding to a bnf that has not been prescribed in the last 13 months.
IC_{Δ}	Strength	The TPD Information Component as calculated in [128]
lowerIC_{Δ}	Strength	The lower 95% interval of the Information Component as calculated in [128]
Age STDEV	Specificity	Standard deviation of patient's age who experience medical event after drug divided by standard deviation of the ages for all the patients.
Gender Ratio	Specificity	Male proportion of patients experiencing the medical event within 30 days of the drug divided by male proportion of patients prescribed the drug.
RR drug / RR bnf	Specificity	The RR of the drug divided by the RR for all the drugs in the same family.
Read Code Level	Specificity	The specificity level of the medical event: general (level 1)- specific (level 5).
ABratio Level 2	Temporality	How often the level 2 version of the medical event is recorded after the prescription compared to before.
ABratio Level 3	Temporality	How often the level 3 version of the medical event is recorded after the prescription compared to before.
LEOPARD [161]	Temporality	1 if the drug is prescribed significantly more after the medical event than before, 0 otherwise.
OE_{filt1} [128]	Temporality	1 if the IC_{Δ} is greater the month before the drug than the month after, 0 otherwise.
OE_{filt2} [128]	Temporality	1 if the IC_{Δ} is greater on the day of prescription compared to the month after, 0 otherwise.

Table C.3: Attribute Summary Table

Feature	Criterion	Description
Dosage Ratio	Dosage	Average dosage of patients experiencing the medical event within 30 days of the drug divided by average dosage of patients prescribed the drug.
High Low Ratio	Dosage	Proportion of patients given the highest dosage that experience the medical event (within 30 days) divided by the proportion of patients given the lowest dosage that experience the medical event (within 30 days).
Spearman's rank	Dosage	The Spearman's rank correlation coefficient between the patient dosage and $\{0, 1\}$ indicating if the patient experienced the medical event within 30 days.
Pearson product-moment	Dosage	The Pearson product-moment correlation coefficient between the patient dosage and $\{0, 1\}$ indicating if the patient experienced the medical event within 30 days.
Repeat_1	Experiment	Number of patients that have medical event in at least two distinct hazard periods and not in their non-hazard periods divided by the number of patients that have at least two distinct hazard periods and have medical event in one hazard period.
Repeat_2	Experiment	Number of patients that have medical event in two distinct hazard periods and not in their non-hazard periods divided by the occurrence in the non-hazard periods.

available via the temporal and strength attributes. The experiment and dosage attributes investigated in this paper did not offer sufficient additional information than what could be gained from the RD_{13BNF} or the LEOPARD attributes, although there does appear to be some correlation between the class labels and both the Pearson's correlation rank attribute and the Repeats attributes.

The reason the dosage attributes did not have a greater correlation with the class labels may be due to a limiting factor of comparing different measurement types. The dosages can be recorded via different measurement types for example 'mg', '%', 'mm x cm xcm' or the measure type may be missing. As it is difficult to determine if x quantity of 'mg' is greater than y quantity of '%', the dosage attributes were calculated only considering prescriptions measured in 'mg' (as this was the most popular). Unfortunately this resulted in occasional issues due to 'mg' measured prescriptions of some drugs investigated always being the same quantity or many prescriptions of a drug not being included in the dosage attribute calculations. The experiment attributes were also limited if the drug investigated was rarely repeated. Furthermore, the experiment attributes may have been biased in this study due to using known ADRs, as if an ADR is known and a patient experiences the ADR after the drug then the doctor is likely to notice this and not prescribed the drug to that patient in the future. One possible way to overcome this issue would be to use only newly discovered ADRs in the data as the medical records may be more likely to have patients, who at the time unknowingly experienced the ADRs, having a repeat prescription.

Table C.4: The results of the CFS algorithm ordered by the measure of correlation with the class labels. Attributes not selected by the CFS algorithm have the attribute they are most correlated to listed in the CFS rank column.

Attribute	Class Correlation	CFS Rank
LEOPARD	0.3238	1
OE_{filt1}	0.2637	LEOPARD
OE_{filt2}	0.2618	LEOPARD
RD_{13BNF}	0.2347	2
RD_{13d}	0.2248	LEOPARD
RD	0.2231	RD_{13BNF}
ABratio Lv3	0.2231	3
ABratio Lv2	0.1755	ABratio Lv3
RR_{13d}	0.1593	RD_{13BNF}
OR_{13d}	0.1593	RD_{13BNF}
RR_{13BNF}	0.1514	RD_{13BNF}
OR_{13BNF}	0.1514	RD_{13BNF}
RR	0.1408	RD_{13BNF}
OR	0.1408	RD_{13BNF}
$lowerIC_{\Delta}$	0.135	RD_{13BNF}
Pearson rank	0.1029	RD_{13BNF}
Gender Ratio	0.0663	4
$Repeats_1$	0.0651	LEOPARD
$Repeats_2$	0.0651	LEOPARD
IC_{Δ}	0.0608	RD_{13BNF}
Read Code Lv	0.0279	5
RR_{Drug}/RR_{BNF}	0	-
Dosage Ratio	0	-
High Low Ratio	0	-
Age STDEV	0	-
Spearman's' rank	0	-

Conclusion

In this paper we have applied feature selection to attributes we generated based on the Bradford Hill causality criteria to determine suitable attributes to be used by a general learning algorithm to identify side effects in LODs. This is the first time suitable attributes for identifying causal relations between prescribed drugs and medical events have been explored and the results now present the opportunity to develop novel learning algorithms. We have found that the specificity attributes offer additional information for ADR signalling and it would be advantageous to include them into ADR signalling algorithms. Unfortunately the experiment and dosage attributes were not very correlated with the class labels but this is likely to be due to current limitations.

Possible future work could focus on developing a way to compare prescriptions with different measurement types so all the prescription data can be used for calculating the dosage attributes or involve developing attributes that cover the remaining Bradford Hill causality criteria (plausibility, coherence, consistency and analogy).

Appendix D

SAP Result Tables

ADR Signalling Framework Results

The signals generated by the SAP framework on the unlabelled data for the drugs Nifedipine, Ciprofloxacin, Ibuprofen, Budesonide and Naproxen.

Nifedipine

Read Code	Medical Event	Frequency
N131.	Cervicalgia - pain in neck	3659
D00..	Iron deficiency anaemias	1281
81H..	Dressing of wound	7674
K15..	Cystitis	3156
461..	Urine exam. - general	1482
R090.	[D]Abdominal pain	2520
H00..	Acute nasopharyngitis	1037

1C9..	Sore throat symptom	2749
1D13.	C/O: a pain	5013
16C2.	Backache	2527
M18z.	Pruritus NOS	1976
413..	Laboratory test requested	6064
H05z.	Upper respiratory infection NOS	6865
M0z..	Skin and subcut tissue infection NOS	454
G84..	Haemorrhoids	1494
1972.	Epigastric pain	1537
1M10.	Knee pain	2690
M2yz.	Other skin and subcutaneous tissue disease NOS	2102
A53..	Herpes zoster	1549
M01..	Furuncle - boil	342
K190z	Urinary tract infection, site not specified NOS	5167
M12z1	Eczema NOS	2785
16C6.	Back pain without radiation NOS	2385
2....	Examination / Signs	2845
R021z	[D]Rash and other nonspecific skin eruption NOS	2875
H33..	Asthma	2792
N142.	Pain in lumbar spine	4225
8HQ1.	Refer for X-Ray	2421
H06z0	Chest infection NOS	14291
H27z.	Influenza NOS	678
1C14.	Blocked ear	994

2D82.	O/E - wax in auditory canal	1939
892..	Informed consent for procedure	1976
M03z0	Cellulitis NOS	2326
856..	Acupuncture	597
1625.	Abnormal weight loss	581
M101.	Seborrhoeic dermatitis	803
M2z0.	Skin lesion	931
ZV583	[V]Attention to surgical dressings or sutures	347
ZV681	[V]Issue of repeat prescription	2994
8BMC.	Prescription collected by pharmacy	1206
1922.	Sore mouth	476
H02..	Acute pharyngitis	1055
8C1B.	Nursing care blood sample taken	10786
2516.	Abdomen examined - NAD	902
AB0..	Dermatophytosis including tinea or ringworm	1051
8H5B.	Referred to urologist	975
M0...	Skin and subcutaneous tissue infections	963
8CA..	Patient given advice	5746
E2B..	Depressive disorder NEC	3082
N094K	Arthralgia of hip	2466
1A...	Genitourinary symptoms	1313
41B1.	Blood test due	2474
8H77.	Refer to physiotherapist	1952
F587.	Otalgia	1047

K190.	Urinary tract infection, site not specified	3347
8E...	Physiotherapy/remedial therapy	2474
2F13.	O/E - dry skin	2055
Z4A..	Discussion	4065
AB2..	Candidiasis	1156
TC...	Accidental falls	4253
SP255	Postoperative wound infection, unspecified	568
25Q..	O/E - rectal examination done	484
22L..	O/E - wound healing	325
H26..	Pneumonia due to unspecified organism	382
M180.	Pruritus ani	568
D21z.	Anaemia unspecified	1690
M21z1	Skin tag	474
N2471	Leg cramps	1915
8BAA.	Pain relief	1455
F502z	Otitis externa NOS	1962
1J4..	Suspected UTI	1551
1....	History / symptoms	1610
1C...	Ear/nose/throat symptoms	287
58D..	Ultrasound scan	259
AB01.	Dermatophytosis of nail	939
M07z.	Local infection skin/subcut tissue NOS	986
J43..	Other non-infective inflammatory gastroenteritis and colitis	719

67I..	Advice	1789
22J..	O/E - dead	231
F501.	Infective otitis externa	1395
8B3A1	Medication increased	3450
N094.	Pain in joint - arthralgia	1243
M161z	Psoriasis NOS	599
R062.	[D]Cough	959
176..	C/O - catarrh	370
8B314	Medication review	15307
M111.	Atopic dermatitis/eczema	1320
1C3..	Earache symptoms	383
1CA2.	Hoarse	430
1C12.	Hearing difficulty	704
J520z	Constipation NOS	965
2128.	Patient's condition the same	5198
F1310	Benign essential tremor	91
2227.	O/E - rash present	836
8C9..	Reassurance given	671
A781.	Viral warts	420
J50zz	Intestinal obstruction NOS	134
C2621	Vitamin B12 deficiency	323
8H9..	Planned telephone contact	809
1D14.	C/O: a rash	3294
N143.	Sciatica	2510

M12z0	Dermatitis NOS	1143
K28y6	Epididymal cyst	185
M262.	Sebaceous cyst - wen	659
67E..	Foreign travel advice	1570
73050	Irrigation of external auditory canal for removal of wax	9722
8B21.	Drug prescription	637
2315.	Resp. system examined - NAD	1622
1D15.	C/O: itching	715
4K...	General pathology	1119
85D..	Injection given	737
F51..	Nonsuppurative otitis media + eustachian tube disorders	94
8B3H.	Medication requested	16644
16C5.	C/O - low back pain	1507
H060.	Acute bronchitis	2569
19EA.	Change in bowel habit	812
8P...	Removal of surgical material and sutures	230
M230.	Ingrowing nail	521
E112.	Single major depressive episode	328
R0300	[D]Appetite loss	203
R0040	[D]Dizziness	2585
7G223	Removal of suture from skin NEC	809
J082.	Oral aphthae	645

1C8..	Nasal symptoms OS	301
6896.	Depression screening using questions	12687
16C..	Backache symptom	451
M244.	Folliculitis	402
R021.	[D]Rash and other nonspecific skin eruption	520
7NB16	[SO]Toe NEC	20
19FZ.	Diarrhoea symptom NOS	374
ZV49z	[V]Unspecified limb or other problem	1398
1739.	Shortness of breath	1856
F4E51	Xanthelasma	31
1832.	Ankle swelling	1681
8BAD.	Chemotherapy	236
1982.	Nausea present	440
H17..	Allergic rhinitis	504
M12..	Contact dermatitis and other eczemas	316
2D...	Ear, nose + throat examination	610
5882.	Spirometry	386
M036z	Cellulitis and abscess of leg NOS	534
8H21.	Admit medical emergency unsp.	470
68M..	Spirometry screening	381
70560	Carpal tunnel release	346
8HP2.	Refer for microbiological test	285
Eu32z	[X]Depressive episode, unspecified	832
1AG..	Recurrent urinary tract infections	294

J1544	Helicobacter gastritis	15
ZGB62	Advice about side effects of drug treatment	55
M05..	Impetigo	242
2G5..	O/E - foot	5012
ZGB64	Advice to start drug treatment	112
8B35.	Drug Rx stopped-medical advice	1671
J530.	Anal fissure	173
C3541	Hypercalcaemia NEC	113
N2133	Olecranon bursitis	483
R1057	[D]Glucose, blood level abnormal	318
F4Kz1	Eye pain NOS	522
Z1B13	Change of dressing	348
6A...	Patient reviewed	37926
Eu410	[X]Panic disorder [episodic paroxysmal anxiety]	13
F4D0.	Blepharitis	1266
7K6WS	Arthroscopic acromioplasty	23
165..	Temperature symptoms	217
2FD..	O/E - skin cyst	360
G3111	Unstable angina	120
73130	Myringotomy and insertion of short term grommet	26
1BK..	Worried	329
8B41.	Repeated prescription	6599
N30z8	Bone infection NOS, of other specified site	70

8B3A2	Medication decreased	1337
N0946	Arthralgia of the lower leg	3023
F52z.	Otitis media NOS	624
E2003	Anxiety with depression	790
G57y9	Supraventricular tachycardia NOS	99
M03z1	Abscess NOS	157
G581.	Left ventricular failure	1154
M01z.	Boil NOS	182
8C1..	Nursing care	2782
G57y7	Sinus tachycardia	57
J5730	Rectal haemorrhage	1424
R0350	[D]Excessive thirst	36
ZV700	[V]Routine health checkup	152
8C15.	Nursing care - dressing	1557
G30..	Acute myocardial infarction	1145
R0734	[D]Bloating	107
H041.	Acute tracheitis	541
1BE1.	Problem situation	238
E2C01	Anger reaction	10
168..	Tiredness symptom	1848
J521.	Irritable colon - Irritable bowel syndrome	848
5....	Radiology/physics in medicine	706
F504.	Impacted cerumen (wax in ear)	4067
7N511	[SO]Prostate	57

8HQ2.	Refer for ultrasound investign	403
8B24.	Prescription given no examination of patient	907
SD...	Superficial injury	446
7L143	Intravenous blood transfusion NEC	303
ZV411	[V]Other eye problems	420
AB200	Candidiasis of mouth	445
1B1X.	Behavioural problem	3
ZGB17	Advice to stop treatment	18
8B316	Medication changed	2761
ZGB67	Advice about drug dosage	208
4JK21	High vaginal swab culture negative	7
H170.	Allergic rhinitis due to pollens	768
7L172	Blood withdrawal for testing	12960
7H2B0	Paracentesis abdominis for ascites	6
ZV6D5	[V]Person consulting for explanatn of investigation findings	232
22Q..	Wound observation	237
N2179	Plantar fasciitis	789
R090B	[D]Groin pain	651
M07yz	Other spec local skin/subc infection NOS	646
ZV720	[V]Examination of eyes and vision	114
R1320	[D]Echocardiogram abnormal	28
G83..	Varicose veins of the legs	752
7K6Z2	Injection of therapeutic substance into joint	372

1B5..	Incoordination symptom	2893
AB20.	Candidiasis of mouth and oesophagus	368
79294	Insertion of coronary artery stent	64
A7811	Verruca plantaris	123
A7814	Plain wart	151
F4Kz4	Redness of eye NOS	282
Z1823	Chaperone refused	44
N135z	Torticollis NOS	71
195..	Indigestion symptoms	365
1M...	Pain	634
M12z2	Infected eczema	186
7G2E3	Dressing of skin NEC	731
1BT..	Depressed mood	908
S64..	Intracranial injury NOS	290
196..	Type of GIT pain	437

Table D.1: The medical events signalled by the SAP framework with the random forest classifier for the drug Nifedipine. The medical events are ranked by the confidence returned by the classifier for the medical event belonging to the ADR class.

Ciprofloxacin

Read Code	Medical Event	Frequency
1BT..	Depressed mood	625
2227.	O/E - rash present	329
E2B..	Depressive disorder NEC	779

A53..	Herpes zoster	364
892..	Informed consent for procedure	513
Z4A..	Discussion	2861
66R5.	Rep.presc. treatment changed	515
C04..	Acquired hypothyroidism	324
32...	Electrocardiography	363
D00..	Iron deficiency anaemias	468
R021z	[D]Rash and other nonspecific skin eruption NOS	658
168..	Tiredness symptom	1006
8E...	Physiotherapy/remedial therapy	686
D00y1	Microcytic hypochromic anaemia	164
J082.	Oral aphthae	285
Eu32.	[X]Depressive episode	139
1D14.	C/O: a rash	1310
F4430	Anterior uveitis	9
8C1B.	Nursing care blood sample taken	3263
E2741	Transient insomnia	164
81H..	Dressing of wound	4336
J520z	Constipation NOS	433
R0720	[D]Difficulty in swallowing	90
E200.	Anxiety states	573
R090B	[D]Groin pain	468
J5730	Rectal haemorrhage	465
R0608	[D]Shortness of breath	1143

7L172	Blood withdrawal for testing	3209
R021.	[D]Rash and other nonspecific skin eruption	176
ZV57C	[V]Palliative care	260
ZV682	[V]Expert advice request	43
G5y34	Ventricular hypertrophy	22
G5yy9	Left ventricular systolic dysfunction	11
S5yz1	Muscle injury / strain	46
2127.	Patient's condition worsened	1099
8B311	Medication given	2556
1D13.	C/O: a pain	1931
8H77.	Refer to physiotherapist	602
G580.	Congestive heart failure	542
8B313	Medication commenced	929
Z1B13	Change of dressing	247
Z1K13	Removal of suture from skin	15
Z4G1B	Giving encouragement to continue treatment	7
M18z.	Pruritus NOS	576
8H9..	Planned telephone contact	581
E112.	Single major depressive episode	87
Eu32z	[X]Depressive episode, unspecified	356
ZV681	[V]Issue of repeat prescription	806
8H21.	Admit medical emergency unsp.	340
K2710	Balanitis	77
22C2.	O/E - oedema of ankles	383

R060A	[D]Dyspnoea	397
70652	Nerve conduction studies	18
8BAA.	Pain relief	782
R0300	[D]Appetite loss	88
8C15.	Nursing care - dressing	981
681..	Screening - general	905
771Qz	Diagnostic rigid sigmoidoscopic exam of sigmoid colon NOS	134
ZV583	[V]Attention to surgical dressings or sutures	121
N145.	Backache, unspecified	387
7G2E3	Dressing of skin NEC	351
Ryu8A	[X]Hyperglycaemia, unspecified	22
G84..	Haemorrhoids	487
R0700	[D]Nausea	59
2315.	Resp. system examined - NAD	988
N2470	Swelling of limb	328
AB200	Candidiasis of mouth	501
1B13.	Anxiousness	700
313B.	Audiogram	86
ZZZZZ	converted code	4445
1982.	Nausea present	428
1M10.	Knee pain	685
7C032	Unilateral total orchidectomy - unspecified	39
8C1..	Nursing care	1124

173..	Breathlessness	2023
677B.	Advice about treatment given	1826
M1535	Perioral dermatitis	17
761Fz	Diagnostic fibreoptic endoscopic exam upper GI tract NOS	158
7L17.	Blood withdrawal	2638
70560	Carpal tunnel release	77
8H8..	Follow-up arranged	1439
C3652	Dehydration NEC	90
1B5..	Incoordination symptom	807
423..	Haemoglobin estimation	70
AB220	Candidal balanitis	67
2841.	Confused	434
8B3A3	New medication commenced	293
R0420	[D]Swelling in head or neck	45
7L171	Venesection	610
ZV680	[V]Issue of medical certificate	1019
E2001	Panic disorder	192
ZL233	Under care of district nurse	55
H17..	Allergic rhinitis	158
8B316	Medication changed	727
G581.	Left ventricular failure	318
E2003	Anxiety with depression	254
41D0.	Blood sample taken	1984

R0073	[D]Lethargy	157
8HB2.	Medical follow-up	533
21262	Asthma resolved	106
ZV6D6	[V]Alcohol abuse counselling and surveillance	4
N2174	Achilles tendinitis	130
7L185	Intramuscular injection of vitamin B12	724
7G2A6	Insertion of hormone implant	54
J4101	Ulcerative colitis	103
C3661	Fluid retention	94
S6460	Minor head injury	6
A3B11	Meticillin resistant staphylococcus aureus	73
M1616	Guttate psoriasis	17
7N522	[SO]Epididymis	31
8H4B.	Referred to rheumatologist	132
68...	Screening	724
8H76.	Refer to dietician	150
ZV700	[V]Routine health checkup	48
R1057	[D]Glucose, blood level abnormal	82
H51y7	Malignant pleural effusion	15
R1100	[D]Albuminuria	27
Z174L	Skin care	36
J50zz	Intestinal obstruction NOS	102
C11y3	Impaired fasting glycaemia	45
44120	Urea and electrolytes normal	24

Eu410	[X]Panic disorder [episodic paroxysmal anxiety]	7
F4C71	Subconjunctival haemorrhage	90
A3A0.	Gas gangrene	13
7G2EA	Two layer compression bandage for skin ulcer	31

Table D.2: The medical events signalled by the SAP framework with the random forest classifier for the drug Ciprofloxacin. The medical events are ranked by the confidence returned by the classifier for the medical event belonging to the ADR class.

Ibuprofen

Read Code	Medical Event	Frequency
D00..	Iron deficiency anaemias	1876
198..	Nausea	3084
K190z	Urinary tract infection, site not specified NOS	5945
461..	Urine exam. - general	1842
M28..	Urticaria	1268
81H..	Dressing of wound	10761
2227.	O/E - rash present	1717
D21z.	Anaemia unspecified	2089
16E..	Feels unwell	2129
8H9..	Planned telephone contact	1386
H33..	Asthma	3253
M0z..	Skin and subcut tissue infection NOS	599
K190.	Urinary tract infection, site not specified	4806
H06z0	Chest infection NOS	17499

22L..	O/E - wound healing	600
1B5..	Incoordination symptom	3824
R021z	[D]Rash and other nonspecific skin eruption NOS	3253
SP255	Postoperative wound infection, unspecified	887
Z4A..	Discussion	8370
1D14.	C/O: a rash	6155
1982.	Nausea present	987
168..	Tiredness symptom	2937
535..	Standard chest X-ray	1465
R090.	[D]Abdominal pain	4220
67I..	Advice	3408
1922.	Sore mouth	794
7G2E3	Dressing of skin NEC	817
8HB2.	Medical follow-up	2143
J5730	Rectal haemorrhage	1473
66R5.	Rep.presc. treatment changed	2890
H1y1z	Nasal cavity and sinus disease NOS	907
M0...	Skin and subcutaneous tissue infections	1766
AB2..	Candidiasis	3488
1J4..	Suspected UTI	2814
1A...	Genitourinary symptoms	1976
2315.	Resp. system examined - NAD	2992
C34..	Gout	3709
66R..	Repeat prescription monitoring	3230

H30..	Bronchitis unspecified	1682
G20..	Essential hypertension	7883
4131.	Blood test requested	2484
1C14.	Blocked ear	1257
2F13.	O/E - dry skin	2790
8B314	Medication review	13074
AB200	Candidiasis of mouth	681
E2B..	Depressive disorder NEC	5162
R062.	[D]Cough	1180
J520z	Constipation NOS	1422
M230.	Ingrowing nail	1016
1BT..	Depressed mood	2678
F4D0.	Blepharitis	1298
M07z.	Local infection skin/subcut tissue NOS	1685
Eu32z	[X]Depressive episode, unspecified	2044
G66..	Stroke and cerebrovascular accident unspecified	919
ZV720	[V]Examination of eyes and vision	147
8C1..	Nursing care	4083
8C15.	Nursing care - dressing	2028
1AG..	Recurrent urinary tract infections	500
196..	Type of GIT pain	984
1737.	Wheezing	1553
181..	Palpitations	1860
67E..	Foreign travel advice	3206

182..	Chest pain	10791
413..	Laboratory test requested	7598
1B321	Weakness of leg	177
8C1L.	Wound care	884
23...	Examn. of respiratory system	1535
ZV681	[V]Issue of repeat prescription	3691
81H5.	Change of dressing	635
8B41.	Repeated prescription	6877
R021.	[D]Rash and other nonspecific skin eruption	894
F51y0	Eustachian tube dysfunction	742
F502z	Otitis externa NOS	2593
E112.	Single major depressive episode	569
1Y...	Patient feels well	2684
AB20.	Candidiasis of mouth and oesophagus	513
K28y8	Pain in testis	258
2841.	Confused	1066
D00y1	Microcytic hypochromic anaemia	618
1....	History / symptoms	2793
M05..	Impetigo	1098
6A...	Patient reviewed	70321
7G2E.	Dressing of skin or wound	1255
1739.	Shortness of breath	2489
8CA..	Patient given advice	9574
N2133	Olecranon bursitis	919

73050	Irrigation of external auditory canal for removal of wax	9848
A07y0	Viral gastroenteritis	355
8HQ2.	Refer for ultrasound investign	857
1D13.	C/O: a pain	12907
8C1B.	Nursing care blood sample taken	12573
H02..	Acute pharyngitis	2993
E200.	Anxiety states	2845
AD30.	Scabies	554
8H76.	Refer to dietician	698
N20..	Polymyalgia rheumatica	2441
Z1B13	Change of dressing	498
M03z.	Cellulitis and abscess NOS	1473
G65..	Transient cerebral ischaemia	1212
662..	Cardiac disease monitoring	17235
J0250	Dental abscess	909
1B8..	Eye symptoms	1863
7M0G1	Aspiration of other lesion of organ NOC	65
8H8..	Follow-up arranged	4686
8H5B.	Referred to urologist	1159
J64..	Cholelithiasis	544
7NB00	[SO]Shoulder NEC	60
81HZ.	Wound dressing NOS	1866
R012z	[D]Gait abnormality NOS	155

2D82.	O/E - wax in auditory canal	2078
ZV680	[V]Issue of medical certificate	4916
M18z.	Pruritus NOS	2310
892..	Informed consent for procedure	3444
TGyz3	Accidental injury NOS	70
D41yz	Other specified disease of blood or blood forming organ NOS	148
1D131	C/O - pain in hallux	256
S2420	Fracture of scaphoid	83
R0701	[D]Vomiting	170
H060.	Acute bronchitis	3311
G33..	Angina pectoris	3241
R0222	[D]Lump, localized and superficial	517
212..	Patient examined	3042
7G2B1	Injection of therapeutic substance into skin	183
41D0.	Blood sample taken	8004
2128.	Patient's condition the same	14124
N30z8	Bone infection NOS, of other specified site	275
G30..	Acute myocardial infarction	629
4K...	General pathology	1181
8BAA.	Pain relief	4430
7G22.	Removal of repair material from skin	1713
F501.	Infective otitis externa	2790
AB0..	Dermatophytosis including tinea or ringworm	1382

173B.	Nocturnal cough / wheeze	492
1C9..	Sore throat symptom	8963
N2243	Ganglion unspecified	216
8C9..	Reassurance given	1597
2126.	Patient's condition improved	22539
M15y1	Intertrigo	1556
M0203	Paronychia of finger	249
Z1B..	Dressing of skin or wound	677
7L172	Blood withdrawal for testing	13906
AB220	Candidal balanitis	102
A53..	Herpes zoster	2338
C3652	Dehydration NEC	117
8H7R.	Refer to chiropodist	941
7G251	Drainage of lesion of skin NEC	189
H00..	Acute nasopharyngitis	1627
ZV49z	[V]Unspecified limb or other problem	3605
F1382	Spasmodic torticollis	118
ZL146	Under care of deputising GP	245
R0052	[D]Insomnia NOS	2374
246..	O/E - blood pressure reading	7872
7G0C1	Biopsy of lesion of skin NEC	35
N2410	Myalgia unspecified	2246
SN52.	Drug hypersensitivity NOS	251
M03z0	Cellulitis NOS	3272

8HQ1.	Refer for X-Ray	7376
F4Kz4	Redness of eye NOS	282
ZV583	[V]Attention to surgical dressings or sutures	550
R0400	[D]Facial pain	568
7L171	Venesection	1946
7L11y	Other specified injection of therapeutic substance	66
173..	Breathlessness	3310
M101.	Seborrhoeic dermatitis	1018
R065z	[D]Chest pain NOS	195
H26..	Pneumonia due to unspecified organism	489
19FZ.	Diarrhoea symptom NOS	728
677B.	Advice about treatment given	6676
8B311	Medication given	8307
19B..	Flatulence/wind	629
32...	Electrocardiography	2209
SN530	Allergic reaction	110
ZV6D5	[V]Person consulting for explanatn of investigation findings	283
7L17.	Blood withdrawal	11641
1B320	Weakness of arm	68
M244.	Folliculitis	773
R0608	[D]Shortness of breath	2053
F4G01	Orbital cellulitis	56

N23y4	Spasm of muscle	597
E2001	Panic disorder	933
1954.	Indigestion	1326
4617.	MSU = abnormal	173
ZGB62	Advice about side effects of drug treatment	75
SP2y2	Postoperative pain	289
1D15.	C/O: itching	1097
D00zz	Iron deficiency anaemia NOS	52
N2241	Ganglion of joint	115
R082.	[D]Retention of urine	440
N2457	Shoulder pain	1446
R0043	[D]Vertigo NOS	1898
C2943	Iron deficiency	130
ZZZZZ	Converted code	21209
1A7..	Vaginal discharge symptom	1047
585..	Other diagnostic ultrasound	412
4618.	Urine dipstick test	1090
704A0	Therapeutic lumbar epidural injection	121
M12..	Contact dermatitis and other eczemas	580
F586.	Otorrhoea	286
7G090	Cauterisation of lesion of skin NEC	230
M200z	Corns NOS	127

Table D.3: The medical events signalled by the SAP framework with the random forest classifier for the drug Ibuprofen. The medical events are ranked by the confidence returned by the classifier for the medical event belonging to the ADR class.

Budesonide

Read Code	Medical Event	Frequency
R090.	[D]Abdominal pain	1436
K190z	Urinary tract infection, site not specified NOS	1626
N245.	Pain in limb	5360
1A1..	Micturition frequency	892
892..	Informed consent for procedure	1286
D00..	Iron deficiency anaemias	670
413..	Laboratory test requested	3881
8C9..	Reassurance given	540
19C..	Constipation	2650
A53..	Herpes zoster	703
1D14.	C/O: a rash	2553
2227.	O/E - rash present	601
N142.	Pain in lumbar spine	2397
1B8..	Eye symptoms	875
1B8Z.	Eye symptom NOS	352
1M10.	Knee pain	1999
N131.	Cervicalgia - pain in neck	2365
Z4A..	Discussion	3827
K190.	Urinary tract infection, site not specified	1513
TC...	Accidental falls	1528
R021z	[D]Rash and other nonspecific skin eruption NOS	1396
16C2.	Backache	1185

M244.	Folliculitis	394
1A...	Genitourinary symptoms	760
AB0..	Dermatophytosis including tinea or ringworm	645
M03z0	Cellulitis NOS	1236
16C5.	C/O - low back pain	1189
461..	Urine exam. - general	518
K15..	Cystitis	1687
2F13.	O/E - dry skin	1066
F501.	Infective otitis externa	1082
M0z..	Skin and subcut tissue infection NOS	272
1B5..	Incoordination symptom	1407
8H77.	Refer to physiotherapist	1176
8B24.	Prescription given no examination of patient	524
ZV583	[V]Attention to surgical dressings or sutures	233
M12z0	Dermatitis NOS	514
16C6.	Back pain without radiation NOS	1831
81H..	Dressing of wound	3827
M07z.	Local infection skin/subcut tissue NOS	736
1BT..	Depressed mood	1238
1J4..	Suspected UTI	1133
41D0.	Blood sample taken	3193
8B4..	Previous treatment continue	5212
D21z.	Anaemia unspecified	561
1D15.	C/O: itching	494

2D82.	O/E - wax in auditory canal	762
8H9..	Planned telephone contact	664
E2001	Panic disorder	432
7L172	Blood withdrawal for testing	6199
8C1B.	Nursing care blood sample taken	4453
M101.	Seborrhoeic dermatitis	438
N143.	Sciatica	1198
7L17.	Blood withdrawal	4396
M02z.	Cellulitis and abscess of digit NOS	254
6A5..	Ongoing review	273
8B41.	Repeated prescription	3431
22J..	O/E - dead	125
N0946	Arthralgia of the lower leg	1112
J155.	Gastritis unspecified	324
8CA..	Patient given advice	4064
N094K	Arthralgia of hip	1182
66R..	Repeat prescription monitoring	1212
E2B..	Depressive disorder NEC	1434
AB01.	Dermatophytosis of nail	574
16E..	Feels unwell	789
16C..	Backache symptom	314
M2yz.	Other skin and subcutaneous tissue disease NOS	986
E200.	Anxiety states	1170
J520z	Constipation NOS	471

6A...	Patient reviewed	28334
2516.	Abdomen examined - NAD	557
677B.	Advice about treatment given	2990
2126.	Patient's condition improved	8208
H06z0	Chest infection NOS	20537
1955.	Heartburn	634
8C1..	Nursing care	1349
7G22.	Removal of repair material from skin	668
F59..	Hearing loss	653
E2273	Impotence	607
F502z	Otitis externa NOS	783
M0...	Skin and subcutaneous tissue infections	634
G20..	Essential hypertension	2258
N0945	Arthralgia of the pelvic region and thigh	519
M12..	Contact dermatitis and other eczemas	223
1C3..	Earache symptoms	556
4618.	Urine dipstick test	355
R0300	[D]Appetite loss	82
M111.	Atopic dermatitis/eczema	1739
R090B	[D]Groin pain	358
2....	Examination / Signs	3388
M12z1	Eczema NOS	2033
424..	Full blood count - FBC	451
1C9..	Sore throat symptom	3604

8C15.	Nursing care - dressing	771
7....	Operations, procedures, sites	259
ZV725	[V]Radiological examination NEC	63
22L..	O/E - wound healing	182
8E...	Physiotherapy/remedial therapy	1413
58D..	Ultrasound scan	205
8BAA.	Pain relief	890
1....	History / symptoms	1628
8BI..	Other medication review	448
M2z0.	Skin lesion	454
22C2.	O/E - oedema of ankles	577
M12z2	Infected eczema	276
J0931	Coated tongue	33
R0734	[D]Bloating	71
16Z3.	Recurrence of problem	499
Z1B..	Dressing of skin or wound	252
N2132	Lateral epicondylitis of the elbow	570
J5730	Rectal haemorrhage	516
J521.	Irritable colon - Irritable bowel syndrome	787
4131.	Blood test requested	1374
73050	Irrigation of external auditory canal for removal of wax	3217
F301z	Trigeminal neuralgia NOS	65
2128.	Patient's condition the same	3380

662..	Cardiac disease monitoring	6147
K3110	Gynaecomastia	40
SP255	Postoperative wound infection, unspecified	244
M161z	Psoriasis NOS	265
1C14.	Blocked ear	523
1AA..	Prostatism	298
N145.	Backache, unspecified	682
6896.	Depression screening using questions	5625
6A2..	Coronary heart disease annual review	758
K20..	Benign prostatic hypertrophy	390
N094.	Pain in joint - arthralgia	764
R0902	[D]Colic NOS	37
M01..	Furuncle - boil	244
C34..	Gout	741
M03z1	Abscess NOS	161
M03z.	Cellulitis and abscess NOS	339
N2470	Swelling of limb	362
TJ...	Drugs and other substances-adverse effects in theraputic use	281
N135z	Torticollis NOS	78
19FZ.	Diarrhoea symptom NOS	250
1A53.	Lumbar ache - renal	405
F4Kz1	Eye pain NOS	207
7K36.	Diagnostic arthroscopy of knee	129

8B21.	Drug prescription	309
K271.	Balanoposthitis	98
J0854	Angular stomatitis and cheilitis	170
8B314	Medication review	6558
G65..	Transient cerebral ischaemia	325
2G5..	O/E - foot	1275
196..	Type of GIT pain	369
R0081	[D]Excessive sweating	80
36150	Gastroscopy abnormal	22
AB200	Candidiasis of mouth	1042
33C..	Circulatory function tests	274
67E..	Foreign travel advice	1578
32...	Electrocardiography	789
4....	Laboratory procedures	521
M180.	Pruritus ani	308
F51y0	Eustachian tube dysfunction	388
R021.	[D]Rash and other nonspecific skin eruption	364
M2400	Alopecia unspecified	100
246..	O/E - blood pressure reading	1966
J64..	Cholelithiasis	249
16ZZ.	General symptom NOS	377
F340.	Carpal tunnel syndrome	399
2D...	Ear, nose + throat examination	659
41B1.	Blood test due	1127

7M07z	Cryotherapy to organ NOC NOS	184
46...	Urine examination	294
R0103	[D]Tremor NOS	201
SE...	Contusion (bruise) with intact skin	223
K4211	Vulvitis unspecified	88
1D131	C/O - pain in hallux	70
N2133	Olecranon bursitis	256
E2003	Anxiety with depression	528
212..	Patient examined	1243
N2410	Myalgia unspecified	527
AD30.	Scabies	197
1B321	Weakness of leg	44
J573.	Haemorrhage of rectum and anus	231
J0250	Dental abscess	322
1M11.	Foot pain	349
8H4B.	Referred to rheumatologist	239
8H5B.	Referred to urologist	430
K28y8	Pain in testis	70
K10y0	Pyelonephritis unspecified	23
N0940	Arthralgia of unspecified site	255
15C..	Vaginal irritation	135
F504.	Impacted cerumen (wax in ear)	1240
M05..	Impetigo	452
C2943	Iron deficiency	76

ZV700	[V]Routine health checkup	64
8H8..	Follow-up arranged	1847
C3652	Dehydration NEC	28
F52z.	Otitis media NOS	1426
8CAK.	Patient given telephone advice out of hours	570
R0904	[D]Abdominal cramps	43
893..	Post operative monitoring	130
J10y4	Oesopheal reflux without mention of oesophagitis	1108
535..	Standard chest X-ray	914
1J...	Suspected condition	651
Z1B13	Change of dressing	191
ZV411	[V]Other eye problems	204
D00y1	Microcytic hypochromic anaemia	233
F4D0.	Blepharitis	499
G3...	Ischaemic heart disease	781
8A...	Monitoring of patient	199
2127.	Patient's condition worsened	927
N2452	Pain in leg	656

Table D.4: The medical events signalled by the SAP framework with the random forest classifier for the drug Budesonide. The medical events are ranked by the confidence returned by the classifier for the medical event belonging to the ADR class.

Naproxen

Read Code	Medical Event	Frequency
E2B..	Depressive disorder NEC	1677
R090.	[D]Abdominal pain	1077
D00..	Iron deficiency anaemias	627
K190z	Urinary tract infection, site not specified NOS	1507
1C9..	Sore throat symptom	1707
R021z	[D]Rash and other nonspecific skin eruption NOS	1034
1BT..	Depressed mood	858
535..	Standard chest X-ray	402
H06z0	Chest infection NOS	4417
G84..	Haemorrhoids	541
C04..	Acquired hypothyroidism	588
H02..	Acute pharyngitis	640
8H9..	Planned telephone contact	392
D21z.	Anaemia unspecified	797
413..	Laboratory test requested	3664
461..	Urine exam. - general	445
892..	Informed consent for procedure	1173
Z4A..	Discussion	2371
H05z.	Upper respiratory infection NOS	2634
1A...	Genitourinary symptoms	530
81H..	Dressing of wound	2969
M230.	Ingrowing nail	372

1D14.	C/O: a rash	1456
66R..	Repeat prescription monitoring	797
H00..	Acute nasopharyngitis	380
2315.	Resp. system examined - NAD	547
H30..	Bronchitis unspecified	514
K190.	Urinary tract infection, site not specified	902
M07z.	Local infection skin/subcut tissue NOS	574
H060.	Acute bronchitis	823
AB2..	Candidiasis	801
E200.	Anxiety states	707
173..	Breathlessness	873
66R5.	Rep.presc. treatment changed	911
M18z.	Pruritus NOS	638
M244.	Folliculitis	247
2F13.	O/E - dry skin	590
2227.	O/E - rash present	337
7L17.	Blood withdrawal	4328
H1y1z	Nasal cavity and sinus disease NOS	286
1Z...	History/symptom NOS	632
1CA2.	Hoarse	172
16E..	Feels unwell	411
7G22.	Removal of repair material from skin	497
8B4..	Previous treatment continue	2945
662..	Cardiac disease monitoring	5045

R0608	[D]Shortness of breath	556
G20..	Essential hypertension	2348
8H76.	Refer to dietician	199
G66..	Stroke and cerebrovascular accident unspecified	232
H06z1	Lower resp tract infection	867
SP255	Postoperative wound infection, unspecified	278
1982.	Nausea present	232
761Fz	Diagnostic fibreoptic endoscopic exam upper GI tract NOS	146
1B5..	Incoordination symptom	902
8B41.	Repeated prescription	1709
J10y4	Oesopheal reflux without mention of oesophagitis	466
7L172	Blood withdrawal for testing	5052
1737.	Wheezing	395
1922.	Sore mouth	196
8C1..	Nursing care	721
8C17.	Nursing care - injections	388
M05..	Impetigo	226
196..	Type of GIT pain	184
F4D0.	Blepharitis	318
19EA.	Change in bowel habit	234
E2273	Impotence	628
M180.	Pruritus ani	243

1739.	Shortness of breath	571
22L..	O/E - wound healing	147
2516.	Abdomen examined - NAD	326
J1011	Reflux oesophagitis	189
4618.	Urine dipstick test	296
AD30.	Scabies	106
8B314	Medication review	3576
M2300	Ingrowing great toe nail	166
J520z	Constipation NOS	308
ZV681	[V]Issue of repeat prescription	1485
E2001	Panic disorder	231
E2003	Anxiety with depression	440
N2133	Olecranon bursitis	361
7L18.	Intramuscular injection	1066
M12z1	Eczema NOS	764
Z174N	Wound care	63
8C15.	Nursing care - dressing	539
M0...	Skin and subcutaneous tissue infections	488
F4Kz1	Eye pain NOS	160
G3111	Unstable angina	24
R060A	[D]Dyspnoea	233
M2yz.	Other skin and subcutaneous tissue disease NOS	645
6A...	Patient reviewed	21724
D00y1	Microcytic hypochromic anaemia	197

8B3R.	Drug therapy discontinued	534
36140	Gastroscopy normal	21
1J4..	Suspected UTI	702
K28y6	Epididymal cyst	75
23...	Examn. of respiratory system	272
7G2E3	Dressing of skin NEC	263
8BI..	Other medication review	301
1972.	Epigastric pain	800
4....	Laboratory procedures	342
R0701	[D]Vomiting	38
41D0.	Blood sample taken	3007
7C242	Standard circumcision	33
8CA..	Patient given advice	2562
4142.	Blood sample -¿ Haematol Lab	426
1Y...	Patient feels well	765
E112.	Single major depressive episode	107
ZV583	[V]Attention to surgical dressings or sutures	149
G5y34	Ventricular hypertrophy	19
73050	Irrigation of external auditory canal for removal of wax	2577
7L064	Amputation below knee	4
J3030	Unilateral inguinal hernia - simple	56
G30..	Acute myocardial infarction	162
J5747	Anal pain	65

R0700	[D]Nausea	35
K2414	Acute epididymitis	54
8H5B.	Referred to urologist	319
677B.	Advice about treatment given	1684
Eu431	[X]Post - traumatic stress disorder	31
7P051	Ultrasound of abdomen	31
2126.	Patient's condition improved	6857
8CAL.	Smoking cessation advice	1621
ZZZZZ	Converted code	4290
7G2A9	Subcutaneous injection of hormone antagonist	87
G581.	Left ventricular failure	230
M15y1	Intertrigo	441
7H220	Exploratory laparotomy	35
8B311	Medication given	2296
J66y6	Obstructive jaundice NOS	13
8H21.	Admit medical emergency unsp.	115
F587.	Otalgia	510
44121	Urea and electrolytes abnormal	48
R090F	[D]Acute abdomen	12
8B42.	Previous treatment repeat	1333
8B24.	Prescription given no examination of patient	293
1B8Z.	Eye symptom NOS	274
7701z	Other excision of appendix NOS	26
SK160	Other hip injuries	31

M111.	Atopic dermatitis/eczema	444
ZL233	Under care of district nurse	20
2841.	Confused	203
7G2E.	Dressing of skin or wound	324
R062.	[D]Cough	284
G3...	Ischaemic heart disease	641
AB0..	Dermatophytosis including tinea or ringworm	421
Z1779	Outpatient care	30
M161z	Psoriasis NOS	244
32...	Electrocardiography	579
R047.	[D]Epistaxis	226
M271.	Non-pressure ulcer lower limb	980
4131.	Blood test requested	969
R0822	[D]Acute retention of urine	13
G65..	Transient cerebral ischaemia	282
79360	Implantation of intravenous cardiac pacemaker system	15
8H7R.	Refer to chiropodist	260
1968.	Abdominal discomfort	185
ZV680	[V]Issue of medical certificate	1629
R0905	[D]Epigastric pain	52
E2900	Grief reaction	65
22C2.	O/E - oedema of ankles	434
2G5..	O/E - foot	920

1B321	Weakness of leg	33
M270.	Decubitus (pressure) ulcer	161
7L185	Intramuscular injection of vitamin B12	686
R0733	[D]Abdominal distension, gaseous	19
Z1B13	Change of dressing	166
173B.	Nocturnal cough / wheeze	94
7NB13	[SO]Lower leg NEC	98
7G2EA	Two layer compression bandage for skin ulcer	14
R021.	[D]Rash and other nonspecific skin eruption	172
M0z..	Skin and subcut tissue infection NOS	149
7G251	Drainage of lesion of skin NEC	58
G3071	Acute non-ST segment elevation myocardial infarction	27
8BAA.	Pain relief	1332
F4504	Ocular hypertension	147
22Q..	Wound observation	164
7B2A.	Diagnostic cystoscopy	293
A3Ay2	Clostridium difficile infection	17
K3110	Gynaecomastia	30
8C1L.	Wound care	190
2128.	Patient's condition the same	3954
7G2E1	Dressing of burnt skin NEC	44
H03..	Acute tonsillitis	544
8HB2.	Medical follow-up	851

J50zz	Intestinal obstruction NOS	46
7L1H0	Direct current cardioversion	10
7L123	Myocardial perfusion scan	13
J5031	Faecal impaction	16
N2243	Ganglion unspecified	74
8HQ1.	Refer for X-Ray	2396
7G033	Excision of lesion of skin NEC	146
Eu32z	[X]Depressive episode, unspecified	488
N094M	Arthralgia of knee	322
81HZ.	Wound dressing NOS	497
M1610	Psoriasis unspecified	416
7B2Az	Diagnostic cystoscopy NOS	115
F4E51	Xanthelasma	28
ZGB66	Advice to stop drug treatment	35
A0745	Helicobacter pylori gastrointestinal tract infection	10
7G2B1	Injection of therapeutic substance into skin	17
K10y0	Pyelonephritis unspecified	14
G831.	Varicose veins of the leg with eczema	315
1....	History / symptoms	695
761F1	Diagnostic gastroscopy NEC	74
44120	Urea and electrolytes normal	68
7G223	Removal of suture from skin NEC	363

782Gz	Diagnostic endosc retrograde exam bile+pancreatic ducts NOS	15
77352	Injection of sclerosing substance into haemorrhoid	29
8B35.	Drug Rx stopped-medical advice	529
M07z1	Infection toe	95
F5611	Benign paroxysmal positional vertigo or nystagmus	232
8CA40	Pt advised re wt reducing diet	171
SK150	Other finger injuries, unspecified	89
Ryu8A	[X]Hyperglycaemia, unspecified	7
Eu411	[X]Generalized anxiety disorder	15
C3540	Hypocalcaemia NEC	16
R0102	[D]Spasms NOS	14
F4005	Eye infection	11
1C14.	Blocked ear	325
R082.	[D]Retention of urine	103
R0400	[D]Facial pain	141
R1100	[D]Albuminuria	14
G2...	Hypertensive disease	1284
F4200	Background diabetic retinopathy	78
C11y3	Impaired fasting glycaemia	96
C2621	Vitamin B12 deficiency	103
G57y7	Sinus tachycardia	12

8C9..	Reassurance given	291
761Fy	Diagnostic fibreoptic endoscopic exam upper GI tract OS	15
R0043	[D]Vertigo NOS	534
Z4G1B	Giving encouragement to continue treatment	23
7M05z	Laser therapy to organ NOC NOS	11
8B3A1	Medication increased	956
Z1K13	Removal of suture from skin	22
ZV57C	[V]Palliative care	84
SD...	Superficial injury	202
TE640	Insect bite NOS	130
246..	O/E - blood pressure reading	2878
M0212	Paronychia of toe	62
8BAB.	Pain control	461
78105	Endoscopic cholecystectomy	63
42QE0	INR - international normal ratio normal	3
7M371	Radiotherapy NEC	145
SN52.	Drug hypersensitivity NOS	78
19FZ.	Diarrhoea symptom NOS	140
2D82.	O/E - wax in auditory canal	472
K253.	Phimosis	37
G73z0	Intermittent claudication	209
4K1..	Histology	149
8A...	Monitoring of patient	91

1AG..	Recurrent urinary tract infections	107
2127.	Patient's condition worsened	861
R090z	[D]Abdominal pain NOS	39
ZL146	Under care of deputising GP	34
R0904	[D]Abdominal cramps	27
68...	Screening	1200
N2470	Swelling of limb	342
ZV49z	[V]Unspecified limb or other problem	1085
ZV654	[V]Other counselling NEC	93
M03z0	Cellulitis NOS	828
7B2B5	Insertion of urethral catheter	4
R1431	[D]Electrocardiogram (ECG) abnormal	12
K20..	Benign prostatic hypertrophy	390
31340	Audiogram bilateral abnormality	7
1B13.	Anxiousness	582
1C8..	Nasal symptoms OS	166
77282	Examination of rectum under anaesthetic	11
R0901	[D]Abdominal colic	79
72550	Trabeculectomy	29
F563.	Labyrinthitis	193
TJ...	Drugs and other substances-adverse effects in theraputic use	422
H33zz	Asthma NOS	26
R0720	[D]Difficulty in swallowing	63

6791.	Health ed. - smoking	2170
M2y45	Epidermal cyst	35
F501.	Infective otitis externa	600
7NB16	[SO]Toe NEC	20
NyuBC	[X]Osteopenia	132
G5730	Atrial fibrillation	283
G580.	Congestive heart failure	368
7M0G1	Aspiration of other lesion of organ NOC	20
Z1745	Ear care	19
M01..	Furuncle - boil	225
679..	Health education - subject	188
F4Ey4	Cyst of eyelid NOS	20
7717z	Other excision of colon NOS	8
Z174O	Post-surgical wound care	16
N0810	Loose body in joint, unspecified joint	19
K272z	Other penile inflammatory disorder NOS	5
M1271	Sunburn	39
8HB20	Medical follow-up - normal	69
G20z.	Essential hypertension NOS	201
ZV6D5	[V]Person consulting for explanatn of investigation findings	149
7G2AC	Insertion of gonadorelin analogue implant	21
G5731	Atrial flutter	20
7M340	Local anaesthetic nerve block	36

R012z	[D]Gait abnormality NOS	26
AB200	Candidiasis of mouth	181
R1103	[D]Microalbuminuria	41
8BA..	Other misc. therapy	174
7920y	Saphenous vein graft replacement of coronary artery OS	7

Table D.5: The medical events signalled by the SAP framework with the random forest classifier for the drug Naproxen. The medical events are ranked by the confidence returned by the classifier for the medical event belonging to the ADR class.

References

[1] J. Almenoff, J. M. Tonning, A. L. Gould, and et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, 28(11):981–1007, 2005. 14, 21

[2] E. Alpaydin. *Introduction to machine learning*. MIT press, 2004. 53

[3] Y. Alvarez, A. Hidalgo, F. Maignen, and J. Slattery. Validation of statistical signal detection procedures in EudraVigilance post-authorization data. *Drug Safety*, 33(6):475–487, 2010. 4, 28, 50

[4] J. Andrés-Ferrer, D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. On the use of different loss functions in statistical pattern recognition applied to machine translation. *Pattern Recognition Letters*, 29(8):1072–1081, 2008. 58

[5] N. Atias and R. Sharan. An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology*, 18(3):207–218, 2011. 137, 188

[6] M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards

bridging theory and practice. In *Advances in Neural Information Processing Systems*, pages 89–96, 2004. 78

[7] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of* 19^{th} *International Conference on Machine Learning (ICML)*, 2002. 79, 80, 83, 149, 151

[8] A. Bate. Bayesian confidence propagation neural network. *Drug Safety*, 30(7):623–625, 2007. 23

[9] A. Bate and S. J. W. Evans. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18(6):427–436, 2009. 24, 28, 122

[10] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998. 24

[11] J. A. Berlin, S. C. Glasser, and S. S. Ellenberg. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *American Journal of Public Health*, 98(8):1366–1371, 2008. 3

[12] J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West. Bayesian methods in pharmacovigilance. pages 421–438. Oxford University Press, 2011. 23, 29

[13] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and

other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008. 72

[14] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the* 21^{st} *International Conference on Machine Learning (ICML)*, page 11. ACM, 2004. 80

[15] B. T. Blak, M. Thompson, H. Dattani, and A. Bourke. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Informatics in Primary Care*, 19(4):251–255, 2011. 29

[16] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the* 11^{th} *Annual Conference on Computational Learning Theory*, pages 92–100. ACM, 1998. 77

[17] F. T. Bourgeois, M. W. Shannon, C. Valim, and K. D. Mandl. Adverse drug events in the outpatient setting: an 11-year national analysis. *Pharmacoepidemiology and Drug Safety*, 19(19):901–910, 2011. 3

[18] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004. 58

[19] A. Bradford-Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300, 1965. 5, 17, 18, 84, 121, 123, 124, 126, 127, 135, 161

[20] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. 49

[21] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth International Group, 1984. 62

[22] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 69, 70

[23] L. Breiman. Bias, variance, and arcing classifiers. 1996. 69, 70

[24] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 71, 73

[25] T. Brodniewicz and G. Grynkiewicz. Preclinical drug development. *Acta Poloniae Pharmaceutica*, 67(6):578–585, 2010. 2

[26] J. S. Brown, M. Kulldorff, K. A. Chan, R. L. Davis, D. Graham, P. T. Pettus, S. E. Andrade, M. A. Raebel, L. Herrinton, D. Roblin, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety*, 16(12):1275–1284, 2007. 31, 41

[27] R. P. Burns and R. Burns. *Business research methods and statistics using SPSS*. Sage, 2008. 65

[28] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the* 10^{th} *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 69–78, 2004. 97

[29] O. Caster and I. R. Edwards. Reflections on attribute and decision in pharmacoviglance. *Drug Safety*, 33(10):805–809, 2010. 13

[30] O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006. 54, 74, 75, 82, 84

[31] Z. Clancy, S. W. Keith, C. Rabinowitz, M. Ceccarelli, J. J. Gagne, and V. Maio. Statins and colorectal cancer risk: a longitudinal study. *Cancer Causes & Control*, pages 1–6, 2013. 30

[32] D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 4(1):17–32, 2003. 79

[33] P. Coloma, P. A. P, F. Salvo, M. Schuemie, C. Ferrajolo, A. Pariente, A. Fourrier-Rglat, M. Molokhia, V. Patadia, J. van der Lei, M. Sturkenboom, and G. Trifirò. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Safety*, 31(1):13–23, 2013. 51, 88, 160

[34] P. M. Coloma, M. J. Schuemie, G. Trifirò, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, L. Pedersen, et al. Combining electronic healthcare databases in europe to allow for large-scale drug safety monitoring: the EU-ADR project. *Pharmacoepidemiology and Drug Safety*, 20(1):1–11, 2011. 44

[35] P. M. Coloma, G. Trifirò, M. J. Schuemie, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, G. Picelli, G. Corrao, L. Pedersen, et al. Electronic

healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiology and Drug Safety*, 21(6):611–621, 2012. 15

[36] A. Coppen, M. Abou-Saleh, P. Milln, J. Bailey, and K. Wood. Decreasing lithium dosage reduces morbidity and side-effects during prophylaxis. *Journal of Affective Disorders*, 5(4):353–362, 1983. 127

[37] M. Cord and P. Cunningham. *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008. 55

[38] O. P. Corrigan. A risky business: the detection of adverse drug reactions in clinical trials and post-marketing exercises. *Social Science & Medicine*, 55(3):497–507, 2002. 2, 3

[39] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 60, 83

[40] M. Culp and G. Michailidis. An iterative algorithm for extending learners to a semi-supervised setting. *Journal of Computational and Graphical Statistics*, 17(3):545–571, 2008. 75

[41] S. Dasgupta, M. L. Littman, and D. McAllester. Pac generalization bounds for co-training. *Advances in Neural Information Processing Systems*, 1:375–382, 2002. 78

[42] T. M. Davidson and W. M. Smith. The bradford hill criteria and zinc-induced anosmia: a causality analysis. *Archives of Otolaryngology – Head & Neck Surgery*, 136(7):673–676, 2010. 112, 138

[43] E. C. Davies, C. F. Green, S. Taylor, P. R. Williamson, D. R. Mottram, and M. Pirmohamed. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *Public Library of Science one*, 4(2):e4439, 2009. 3, 4

[44] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988. 98

[45] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer, 1996. 72

[46] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000. 70, 71

[47] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997. 65

[48] A. K. Dubey, V. M. Castro, W. Mahamaneerat, S. Goryachev, T. D. Wang, C. D. Herrick, V. S. Gainer, and S. N. Murphy. Investigating OMOP false positives with i2b2. http://omop.org/sites/default/files/Investigating%20OMOP%20False%20Positives%20with%20i2b2%20Final%20Report%20November%202012.pdf. Access: 2012-12-20. 87

[49] S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, 6(4):325–327, 1976. 69

[50] W. DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999. 24, 26

[51] S. Džeroski and B. Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3):255–273, 2004. 69

[52] I. R. Edwards. Considerations on causality in pharmacovigilance. *The International Journal of Risk & Safety in Medicine*, 24(1):41–54, 2012. 112

[53] I. R. Edwards and C. Biriell. Harmonisation in pharmacovigilance. *Drug Safety*, 10(2):93–102, 1994. 14

[54] C. P. Farrington. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, 51(1):228–235, 1995. 46

[55] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo. Predicting sample size required for classification performance. *BioMed Central Medical Informatics and Decision Making*, 12(1):8, 2012. 148

[56] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989. 60, 83

[57] P. B. Fontanarosa, D. Rennie, and C. D. DeAngelis. Postmarketing surveillance– lack of vigilance, lack of trust. *The Journal of the American Medical Association*, 292(21):2647–2650, 2004. 4

[58] D. Freedman. From association to causation via regression. *Advances in Applied Mathematics*, 18(1):59 – 110, 1997. 112

[59] Y. Freund. An adaptive version of the boost by majority algorithm. *Machine Learning*, 43(3):293–318, 2001. 71

[60] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Japanese Society For Artificial Intelligence*, 14(5):771–780, 1999. 70

[61] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, pages 23–37. Springer, 1995. 70

[62] R. Gallagher, K. Bird, J. Mason, M. Peak, P. Williamson, A. Nunn, M. Turner, M. Pirmohamed, and R. Smyth. Adverse drug reactions causing admission to a paediatric hospital: a pilot study. *Journal of Clinical Pharmacy and Therapeutics*, 36(2):194–199, 2011. 3

[63] S. A. Goldman. Limitations and strengths of spontaneous reports data. *Clinical Therapeutics*, 20:C40–C44, 1998. 16, 20

[64] K. J. Goodman and C. V. Phillips. *Hill's criteria of causation*. Wiley Online Library, 2005. 127

[65] R. L. Grant, V. M. Drennan, G. Rait, I. Petersen, and S. Iliffe. First diagnosis and management of incontinence in older people with and without dementia in primary care: A cohort study using The Health Improvement Network primary care database. *Public Library of Science Medicine*, 10(8):e1001505, 2013. 29

[66] S. Greenland and H. Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22(1):189–212, 2001. 32

[67] D. A. Grimes and K. F. Schulz. Bias and causal associations in observational research. *The Lancet*, 359(9302):248–252, 2002. 112, 124

[68] K. Haerian, D. Varn, S. Vaidya, L. Ena, H. Chase, and C. Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology & Therapeutics*, 92(2):228–234, 2012. 111

[69] M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. 133, 213

[70] R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, and N. H. Shah. Performance of pharmacovigilance signal-detection algorithms for the FDA Adverse Event Reporting System. *Clinical Pharmacology & Therapeutics*, 93(6):539–546, 2013. 29, 88, 180

[71] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021, 2012. 30, 45

[72] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman. Mining electronic health records for adverse drug effects using regression based methods. pages 100–107, 2010. 37

[73] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. 79

[74] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001. 53

[75] M. Hauben and A. Bate. Data mining in pharmacovigilance. *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery*, 6:341, 2009. 24

[76] M. Hauben, D. Madigan, C. M. Gerrits, L. Walsh, and E. P. Van Puijenbroek. The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety*, 4(5):929–948, 2005. 23

[77] M. Hauben, L. Reich, and S. Chung. Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. *European Journal of Clinical Pharmacology*, 60(10):747–750, 2004. 20

[78] K. Haynes, W. B. Bilker, T. R. TenHave, B. L. Strom, and J. D. Lewis. Temporal and within practice variability in The Health Improvement Network. *Pharmacoepidemiology and Drug Safety*, 20(9):948–955, 2011. 30, 52

[79] L. Hazell and S. A. Shakir. Under-reporting of adverse drug reactions. *Drug Safety*, 29(5):385–396, 2006. 16, 21, 50

[80] P. HealthCare, S. N. Murphy, V. Castro, J. Colecchi, A. Dubey, V. Gainer, C. Herrick, and M. Sordo. Partners healthcare OMOP study report. http://omop.fnih.org/sites/default/files/PHCS_Final_OMOP_StudyReport.pdf, 2011. Accessed: 2013-01-18. 46, 47, 48, 108, 161, 185

[81] E. G. Henrichon Jr and K.-S. Fu. A nonparametric partitioning procedure for pattern classification. *IEEE Transactions on Computers*, 100(7):614–624, 1969. 60, 83

[82] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. 69

[83] M. N. Hocine, P. Musonda, N. J. Andrews, and C. Paddy Farrington. Sequential case series analysis for pharmacovigilance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):213–236, 2009. 31, 42

[84] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. Wiley. com, 2013. 60, 83

[85] R. Hubbard, S. Lewis, J. West, C. Smith, C. Godfrey, L. Smeeth, P. Farrington, and J. Britton. Bupropion and the risk of sudden death: a self-controlled case-series analysis using The Health Improvement Network. *Thorax*, 60(10):848–850, 2005. 29

[86] D. Hume. *A treatise of human nature. Second edition*. Oxford: Oxford University Press, 1978. 123

[87] T. Jaakkola and M. Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Proceedings of the* 6^{th} *International Workshop on Artificial Intelligence and Statistics*. Citeseer, 1997. 48

[88] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999. 79

[89] Y. Ji, H. Ying, J. Tran, P. Dews, A. Mansour, and R. Massanari. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 2012. 41

[90] H. Jin, J. Chen, H. He, C. Kelman, D. McAullay, and C. M. O'Keefe. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):839–853, 2010. 31, 38, 103

[91] H. Jin, J. Chen, C. Kelman, H. He, D. McAullay, and C. M. OKeefe. Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases. In *Proceedings of the* 10^{th} *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 867–876. Springer, 2006. 38, 103, 111

[92] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. pages 841–848, 2002. 57

[93] R. A. Kamtane and V. Jayawardhani. Knowledge, attitude and perception of physicians towards adverse drug reaction (ADR) reporting: A pharmacoepidemiological study. *Asian Journal of Pharmaceutical and Clinical Research*, 5(3):210–214, 2012. 16

[94] J. C. Kando, K. A. Yonkers, and J. O. Cole. Gender as a risk factor for adverse events to medications. *Drugs*, 50(1):1–6, 1995. 88

[95] M. J. Kearns and Y. Mansour. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In *Proceeding of the* 15^{th} *International Conference on Machine Learning (ICML)*, volume 98, pages 269–277. Citeseer, 1998. 61

[96] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006. 131

[97] M. Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. 211

[98] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1):343–349, 2010. 93, 137

[99] B. Kulis. Metric learning: A survey. *Machine Learning*, 5(4):287–364, 2012. 80

[100] R. Kunz and A. D. Oxman. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal*, 317(7167):1185, 1998. 43

[101] P. Langley, W. Iba, and K. Thompson. An analysis of bayesian classifiers. In *Proceeding of the* 10^{th} *National Conference on Artificial Intelligence (AAAI)*, volume 90, pages 223–228, 1992. 60, 83

[102] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Proceeding of the* 19^{th} *IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 87–94. IEEE, 2006. 57

[103] J. Lazarou, B. H. Pomeranz, and P. N. Corey. Incidence of adverse drug reactions in hospitalized patients. *The Journal of the American Medical Association*, 279(15):1200–1205, 1998. 3

[104] J. D. Lewis, W. B. Bilker, R. B. Weinstein, and B. L. Strom. The relationship between time since registration and measured incidence rates in the general practice research database. *Pharmacoepidemiology and Drug Safety*, 14(7):443–451, 2005. 30, 113

[105] J. D. Lewis, R. Schinnar, W. B. Bilker, X. Wang, and B. L. Strom. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and Drug Safety*, 16(4):393–401, 2007. 6, 29

[106] J. Lexchin. Is there still a role for spontaneous reporting of adverse drug reactions? *Canadian Medical Association Journal*, 174(2):191–192, 2006. 19

[107] X. Li, S. Hui, P. Ryan, M. Rosenman, and M. Overhage. Statistical visualization for assessing performance of methods for safety surveillance using electronic databases. *Pharmacoepidemiology and Drug Safety*, 22(5):503–509, 2013. 45

[108] Y. Li, C. Guan, H. Li, and Z. Chin. A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9):1285–1294, 2008. 84

[109] H. Liang-Chin, W. Xiaogang, and Y. C. Jake. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics*, 13(2):313–324, 2013. 138

[110] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. 73

[111] M. Lindquist, I. R. Edwards, A. Bate, H. Fucik, A. M. Nunes, and M. Ståhl. From association to alerta revised approach to international signal analysis. *Pharmacoepidemiology and Drug Safety*, 8(S1):S15–S25, 1999. 14

[112] M. O. Little and A. Morley. Reducing polypharmacy: Evidence from a simple quality improvement initiative. *Journal of the American Medical Directors Association*, 14(3):152–156, 2013. 4

[113] M. Liu, M. E. Matheny, Y. Hu, and H. Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012. 15, 53, 136, 138, 188, 190

[114] Z. Lu. Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. *Drug, Healthcare and Patient Safety*, 1:35–45, 2009. 111

[115] D. Madigan and P. Ryan. Commentary: What can we really learn grom observational studies?: The need for empirical assessment of methodology for active drug safety surveillance and comparative effectiveness research. *Epidemiology*, 22(5):629–631, 2011. 5, 45

[116] O. Z. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*. Springer, 2005. 61, 62, 63

[117] G. Maldonado and S. Greenland. Estimating causal effects. *International Journal of Epidemiology*, 31(2):422–429, 2002. 111

[118] D. J. McLernon, C. M. Bond, P. C. Hannaford, M. C. Watson, A. J. Lee, L. Hazell, and A. Avery. Adverse drug reaction reporting in the uk. *Drug Safety*, 33(9):775–788, 2010. 20

[119] J.-L. Montastruc, A. Sommet, H. Bagheri, and M. Lapeyre-Mestre. Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *British Journal of Clinical Pharmacology*, 72(6):905–908, 2011. 16, 29

[120] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996. 76

[121] N. Moore. The past, present and perhaps future of pharmacovigilance: homage to folke sjoqvist. *European Journal of Clinical Pharmacology*, 69(1):33–41, 2013. 13

[122] S. L. Morgan and C. Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2007. 31

[123] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency

of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006. 58

[124] A. Mullard. Unleashing the mini-sentinel. *Nature Reviews Drug Discovery*, 11(4):255–257, 2012. 44

[125] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001. 68

[126] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000. 148

[127] R. Nisbet, J. Elder, and G. Miner. *Handbook of Statistical Analysis & Data Mining Applications*. Elsevier Incorporation, 2009. 131, 132

[128] N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361–387, 2010. 31, 33, 35, 91, 103, 111, 121, 124, 214

[129] J. L. Oliveira, P. Lopes, T. Nunes, D. Campos, S. Boyer, E. Ahlberg, E. M. Mulligen, J. A. Kors, B. Singh, L. I. Furlong, et al. The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. *Pharmacoepidemiology and Drug Safety*, 22(5):459–467, 2013. 51

[130] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999. 69, 70

[131] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60, 2012. 5, 43, 45

[132] M. Perrio, S. Voss, and S. A. W. Shakir. Application of the bradford hill criteria to assess the causality of cisapride-induced arrhythmia. *Drug Safety*, 30(4):333–346, 2007. 112

[133] T. S. Peters. Do preclinical testing strategies help predict human hepatotoxic potentials? *Toxicologic Pathology*, 33(1):146–154, 2005. 2

[134] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–238, 1991. 38

[135] M. Pirmohamed, S. James, S. Meakin, C. Green, A. K. Scott, T. J. Walley, K. Farrar, B. K. Park, and A. M. Breckenridge. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *British Medical Journal*, 329(7456):15–19, 2004. 3

[136] R. Platt and R. Carnahan. The US food and drug administration's mini-sentinel program. *Pharmacoepidemiology and Drug Safety*, 21(S1):1–303, 2012. 15, 44

[137] E. Poluzzi, E. Raschi, U. Moretti, and F. De Ponti. Drug-induced torsades de pointes: data mining of the public version of the fda adverse event reporting system (aers). *Pharmacoepidemiology and Drug Safety*, 18(6):512–518, 2009. 21

[138] J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987. 62

[139] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan Kaufmann, 1993. 61

[140] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991. 75

[141] C. Reich, P. B. Ryan, P. E. Stang, and M. Rocca. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of Biomedical Informatics*, 45(4):689–696, 2012. 7, 45

[142] J. M. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. Gibson, and R. Hubbard. Attributes for causal inference in electronic healthcare databases. 2013. 127

[143] J. M. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. Gibson, and R. Hubbard. Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Computing*, pages 1–17, 2013. 45, 111, 113, 124

[144] M. J. Rho, S. R. Kim, S. H. Park, K. S. Jang, B. J. Park, and I. Y. Choi. Development common data model for adverse drug signal detection based on multi-center EMR systems. In *Proceedings of* 4^{th} *International Conference on Information Science and Applications (ICISA)*, pages 1–7. IEEE, 2013. 44

[145] R. L. Richesson et al. *Pharmacovigilance*, pages 367–387. Springer, 2012. 15

[146] I. Rish. An empirical study of the naive bayes classifier. In *Proceeding of the* 17^{th} *International Joint Conference on Artificial Intelligence (IJCAM) Workshop on Empirical Methods in Artificial Intelligence*, volume 3, pages 41–46, 2001. 65

[147] T. RJ and H. MH. Safety of newly approved drugs: Implications for prescribing. *The Journal of the American Medical Associaiton*, 287(17):2273–2275, 2002. 88

[148] M. A. Robb, J. A. Racoosin, R. E. Sherman, T. P. Gross, R. Ball, M. E. Reichman, K. Midthun, and J. Woodcock. The US food and drug administration's sentinel initiative: Expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety*, 21(S1):9–11, 2012. 111

[149] L. Rokach and O. Maimon. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(4):476–487, 2005. 61

[150] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. 57

[151] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern epidemiology*. Wolters Kluwer Health, 2008. 121, 124

[152] E. Roux, F. Thiessard, A. Fourrier, B. Begaud, P. Tubert-Bitter, et al. Spontaneous reporting system modelling for data mining methods evalu-

ation in pharmacovigilance. In *Preceding of the AIME Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP)*, 2003. 19

[153] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. 112

[154] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978. 112

[155] P. B. Ryan, G. E. Powell, E. N. Pattishall, and K. J. Beach. Performance of screening multiple observational databases for active drug safety surveillance. In *Proceedings of the* 25^{th} *International Conference on Pharmacoepidemiology & Therapeutic Risk Management (ICPE)*, 2009. 46

[156] P. B. Ryan, D. Madigan, P. E. Stang, J. Marc Overhage, J. A. Racoosin, and A. G. Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in Medicine*, 31(30):4401–4415, 2012. 5, 45, 50, 51, 84, 89, 90, 97, 138, 161, 162, 164, 185

[157] Y. Saeys, I. Inza, and P. Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. 141

[158] S. Schneeweiss. Developments in post-marketing comparative effectiveness research. *Clinical Pharmacology & Therapeutics*, 82(2):143–156, 2007. 16

[159] S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A.

Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009. 45

[160] S. Schneeweiss, T. Stürmer, and M. Maclure. Case–crossover and case–time–control designs as alternatives in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 6(S3):S51–S59, 1997. 45

[161] M. J. Schuemie. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, 20(3):292–299, 2011. 30, 111, 124, 214

[162] M. J. Schuemie, P. M. Coloma, H. Straatman, R. Herings, G. Trifirò, J. N. Matthews, D. Prieto-Merino, M. Molokhia, L. Pedersen, R. Gini, F. Innocenti, G. Mazzaglia, G. Picelli, L. Scotti, J. van der Lei, and M. Sturkenboom. Using electronic health care records for drug safety signal detection: A comparative evaluation of statistical methods. *Medical Care*, 50(10):890–897, 2012. 89

[163] S. A. W. Shakir. *Causality and correlation in pharmacovigilance*, pages 329–343. John Wiley & Sons Limited, 2005. 110

[164] S. A. W. Shakir and D. Layton. Causal association in pharmacovigilance and pharmacoepidemiology. *Drug Safety*, 25(6):467–471, 2002. 112, 186

[165] J. C. Sinclair and M. B. Bracken. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology*, 47(8):881–889, 1994. 123

[166] R. Sirriyeh, S. McClean, and V. Robins. *Introducing Patient Safety: Theory, Policy and Practice*, volume 1357, chapter 1, pages 3–20. SAGE Publications, 2012. 50

[167] A. L. Speirs. Thalidomide and congenital abnormalities. *The Lancet*, 279(7224):303–305, 1962. 13

[168] W. P. Stephenson and M. Hauben. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiology and Drug Safety*, 16(4):359–365, 2007. 16

[169] A. Stevenson. *Oxford dictionary of English*. Oxford Reference Online Premium. OUP Oxford, 2010. 17

[170] N. L. Story. Sexual dysfunction resulting from drug side effects. *Journal of Sex Research*, 10(2):132–149, 1974. 2

[171] B. H. Stricker and B. M. Psaty. Detection, verification, and quantification of adverse drug reactions. *British Medical Journal*, 329(7456):44–47, 2004. 191

[172] B. L. Strom. Potential for conflict of interest in the evaluation of suspected adverse drug reactions. *Journal of the American Medical Association*, 292(21):2643–2646, 2004. 4, 20, 23, 28

[173] B. L. Strom and J. L. Carson. Use of automated databases for pharmacoepidemiology research. *Epidemiologic Reviews*, 12(1):87–107, 1990. 21, 51

[174] M. Suling and I. Pigeot. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics*, 4(4):607–640, 2012. 111

[175] M. Susser. What is a cause and how do we know one? a grammar for pragmatic epidemiology. *American Journal of Epidemiology*, 133(7):635–648, 1991. 127

[176] The National Magazine Company Ltd. NetDoctor.co.uk - the UK's leading independent health website. http://www.netdoctor.co.uk. Accessed: 04-02-2012. 9, 93

[177] K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999. 70

[178] M. J. Titulaer, L. McCracken, I. Gabilondo, T. Armangué, C. Glaser, T. Iizuka, L. S. Honig, S. M. Benseler, I. Kawachi, E. Martinez-Hernandez, et al. Treatment and prognostic factors for long-term outcome in patients with anti-NMDA receptor encephalitis: an observational cohort study. *The Lancet Neurology*, 12(2):157–165, 2013. 46

[179] G. Trifirò, A. Pariente, P. Coloma, , J. Kors, G. Polimeni, G. Miremont-Salam, M. Catania, F. Salvo, A. David, N. Moore, A. Caputi, M. Sturkenboom, M. Molokhia, J. Hippisley-Cox, C. Acedo, J. van der Lei, and A. Fourrier-Reglat. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Safety*, 18(12):1176–1184, 2009. 88, 184

[180] G. Trifirò, V. Patadia, M. J. Schuemie, P. M. Coloma, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, L. Scotti, et al. EU-ADR

healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Studies in Health Technology and Informatics*, 166:25–30, 2011. 16, 17

[181] G. Tripepi, K. Jager, F. Dekker, C. Wanner, and C. Zoccali. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. *Kidney International*, 72(7):789–791, 2007. 122

[182] Uppsala Monitoring Centre. Glossery of terms used in pharmacovigilance. http://www.who-umc.org/graphics/15338.pdf, 2000. Accessed: 2013-04-04. 2

[183] G. Upton and I. Cook. *A dictionary of statistics second edition*. Oxford University Press, 2006. 17

[184] US Food and Drug Administration. Center for Drug Evaluation and Research. Adverse events reporting system (AERS). http://www.fda.gov/cder/aers/default.htm. Accessed: 2012-11-23. 20

[185] E. P. van Puijenbroek, A. Bate, H. G. Leufkens, M. Lindquist, R. Orre, and A. C. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, 11(1):3–10, 2002. 24, 26

[186] O. Vanunu and R. Sharan. A propagation-based algorithm for inferring gene-disease associations. pages 54–52, 2008. 137

[187] V. Vapnik. *The nature of statistical learning theory*. springer, 2000. 58, 59

[188] V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000. 67

[189] V. Vapnik, E. Levin, and Y. Le Cun. Measuring the vc-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994. 60

[190] M. Wadman. Experts call for active surveillance of drug safety. *Nature*, 446(7134):358–359, 2007. 51

[191] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the* 18^{th} *International Conference on Machine Learning (ICML)*, volume 1, pages 577–584, 2001. 79

[192] E. K. Wai, D. M. Roffey, P. Bishop, B. K. Kwon, and S. Dagenais. Causal assessment of occupational bending or twisting and low back pain: results of a systematic review. *The Spine Journal*, 10(1):76–88, 2010. 127

[193] S. D. Walter. The partial area under the summary ROC curve. *Statistics in Medicine*, 24(13):2025–2040, 2005. 98

[194] S. D. Walter. Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*, 53(9):931–939, 2000. 121, 123

[195] B. Wang, B. Spencer, C. Ling, and H. Zhang. Semi-supervised self-training for sentence subjectivity classification. In S. Bergler, editor, *Advances in Artificial Intelligence*, volume 5032 of *Lecture Notes in Computer Science*, pages 344–355. Springer Berlin Heidelberg, 2008. 148

[196] J.-X. Wei, M. Li, Y.-H. Sun, Y. Lu, and H.-M. Xu. A novel method for signal detection of adverse drug reactions based on proportional reporting ratios. *Pharmacy World & Science*, 32(5):658–662, 2010. 24, 26

[197] N. S. Weiss. Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13(1):6–8, 2002. 125

[198] B. Wettermark. The intriguing future of pharmacoepidemiology. *European Journal of Clinical Pharmacology*, 69(1):43–51, 2013. 5, 15, 43

[199] D. Wettschereck, D. W. Aha, and T. Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11(1-5):273–314, 1997. 68

[200] H. J. Whitaker, C. Paddy Farrington, B. Spiessens, and P. Musonda. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine*, 25(10):1768–1797, 2006. 48

[201] A. P. White and W. Z. Liu. Technical note: Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329, 1994. 62

[202] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3):404–408, 2013. 191

[203] R. A. Wilke, H. Xu, J. C. Denny, D. M. Roden, R. M. Krauss, C. A. McCarty, R. L. Davis, T. Skaar, J. Lamba, and G. Savova. The emerging role

of electronic medical records in pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 89(3):379–386, 2011. 5

[204] A. M. Wilson, L. Thabane, and A. Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2):127–134, 2004. 138

[205] B. Woolf. On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19(4):251–253, 1955. 25

[206] World Health Organization. Safety monitoring of medical products. guidlines for setting up and running a pharmacovigilance centre. http://apps.who.int/medicinedocs/en/d/Jh2934e/. Accessed: 2013-04-04. 2

[207] T. Y. Wu, M. H. Jen, A. Bottle, M. Molokhia, and et al. Ten-year trends in hospital admission for adverse drug reactions in england 1999-2009. *Journal of the Royal Society of Medicine*, 103(6):239–250, 2010. 4

[208] S. Xiang, F. Nie, and C. Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600–3612, 2008. 80, 83

[209] L. Xie, J. Li, and P. Bourne. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. *Public Library of Science Computational Biology*, 5(5):e1000387, 2009. 137

[210] E. P. Xing, M. I. Jordan, S. Russell, and A. Ng. Distance metric learning

with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002. 80

[211] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13:1–26, 2012. 80, 81, 82, 83, 151

[212] Y. Yoshihiro, P. Edouard, and K. Masaaki. Drug side-effect prediction based on the integration of chemical and biological spaces. *Journal of Chemical Information and Modeling*, 52(12):3284–3292, 2012. 138

[213] C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, 2002. 192

[214] X. Zhou, S. Murugesan, H. Bhullar, Q. Liu, B. Cai, C. Wentworth, and A. Bate. An evaluation of the THIN database in the OMOP common data model for active drug safety surveillance. *Drug Safety*, 36(2):119–134, 2013. 5, 49, 50, 52, 89, 162, 181, 184

[215] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009. 76, 77, 78, 82, 83, 148

[216] I. Zorych, D. Madigan, P. Ryan, and A. Bate. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical Methods in Medical Research*, 22(1):39–56, 2013. 5, 30, 31, 33, 34, 45, 46, 92, 103

[217] I. Zorych, P. Ryan, and D. Madigan. *Observational Medical Outcomes*

Partnership multi-set case-control estimation specification. Foundation for the National Institutes of Health, 2009. 48