A Novel Semisupervised Algorithm for Rare Prescription Side Effect Discovery

Jenna M. Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson, and Richard B. Hubbard

***Abstract*—Drugs are frequently prescribed to patients with the aim of improving each patient's medical state, but an unfortunate consequence of most prescription drugs is the occurrence of undesirable side effects. Side effects that occur in more than one in a thousand patients are likely to be signaled efficiently by current drug surveillance methods, however, these same methods may take decades before generating signals for rarer side effects, risking medical morbidity or mortality in patients prescribed the drug while the rare side effect is undiscovered. In this paper, we propose a novel computational metaanalysis framework for signaling rare side effects that integrates existing methods, knowledge from the web, metric learning, and semisupervised clustering. The novel framework was able to signal many known rare and serious side effects for the selection of drugs investigated, such as tendon rupture when prescribed Ciprofloxacin or Levofloxacin, renal failure with Naproxen and depression associated with Rimonabant. Furthermore, for the majority of the drugs investigated it generated signals for rare side effects at a more stringent signaling threshold than existing methods and shows the potential to become a fundamental part of post marketing surveillance to detect rare side effects.**

***Index Terms*—Adverse drug reaction (ADR), HUNT, longitudinal healthcare, pharmacovigilance, semisupervised, temporal pattern discovery (TPD).**

I. Introduction

NEGATIVE side effects caused by prescribed medication currently present a huge burden for the healthcare service in terms of causing both patient morbidity or mortality and costing large sums of money [1]–[3]. Investigations have shown that the rate of unwanted side effects has been increasing annually [4], [5]. Possible reasons for this are an increase in the number of annual prescriptions due to an aging population or an increase in polypharmacy, when numerous drugs are prescribed at the same time [6]. Although it is common for a patient to develop side effects due to prescribed medication there is currently no efficient means of identifying all the side effects of a drug. When the side effect is detrimental to the patient's quality of life, it is often referred to as an adverse drug event (ADE) and when the drug causing the ADE is known, it is termed an adverse drug reaction (ADR). A study conducted in the U.K. between November 2001 to April 2002 indicated that 6.5% of admissions to hospital were due to ADRs, with the mortality rate for an ADR patient of 2.3% [7]. Interestingly, it was found that more than 70% of these ADRs were potentially avoidable. A more recent study in Brazil suggests ADRs may be the cause of an even higher proportion of hospital admissions for the elderly as it showed that ADRs were the cause of hospitalization for more than 50% of elderly patients [8]. It also highlighted that a significant factor for developing an ADR was polypharmacy [8].

Some obvious ADRs can be discovered during the experimental stages of a drug's development, but the occurrence of an ADR can depend on a magnitude of factors and it is impossible to investigate all the possible situations that may occur when the drug is taken. For example, testing for ADRs that result from polypharmacy would require clinical trials with millions of people to be able to investigate all the different drug combinations and this is not possible. Due to the limitations of clinical trials, rare ADRs, including fatal ones, are in most circumstances not discovered before a drug is marketed [9], [10]. As a consequence, after a drug is approved and available to patients, possible ADRs are investigated during the whole lifetime of a drug by a process known as postmarketing drug surveillance.

Postmarketing surveillance (such as doctors being vigilant and noticing possible drug and illness associations) can identify common ADRs and in general the more common the ADR is, the fewer the number of patients that need to be prescribed the drug before it is discovered. However, ADRs that occur for drugs that are rarely prescribed or rare ADRs may go unnoticed by medical practitioners and may cause morbidity or mortality in patients that could have been prevented with more efficient drug surveillance methods. For example, it took 23 years before there was sufficient evidence that the drug Tamoxifen used to treat breast cancer caused endometrial cancer in about 1 in 6000 patients [11], [12].

Current methods to discover rare ADRs often involve using a spontaneous reporting system (SRS) database that contains a collection of voluntary suspected drug and ADR reports, such as the database containing information from the U.K. yellow card scheme. The algorithms that signal ADRs by mining SRS databases calculate a measure of how disproportionally more often the medical event is reported with a specific drug of interest compared with any drug. The frequently implemented measures of disproportionality involve using standard epidemiology measures [13], estimating the information component using a neural network approach [14] or calculating a modified version

Manuscript received January 1, 2013; revised July 29, 2013; accepted September 5, 2013. Date of publication September 11, 2013; date of current version March 3, 2014.

J. M. Reps, J. M. Garibaldi, U. Aickelin, and D. Soria are with the School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K. (e-mail: psxjr1@nottingham.ac.uk; jonathan.garibaldi@nottingham.ac.uk; uwe.aickelin@nottingham.ac.uk; daniele.soria@nottingham.ac.uk).

J. E. Gibson and R. B. Hubbard are with the School of Community Health Sciences, University of Nottingham, Nottingham NG7 2RD, U.K. (e-mail: jack.gibson@nottingham.ac.uk; richard.hubbard@nottingham.ac.uk).

Digital Object Identifier 10.1109/JBHI.2013.2281505

of the relative risk by applying a Bayesian model [15]. SRS databases combine reports of possible ADRs from a large population enabling the identification of possible ADR signals more efficiently, but they are known to suffer from underreporting [16] and this causes a lag in the time it takes to confidently signal a potential ADR. The underreporting may also prevent the detection of rare ADRs, as these ADRs may never be suspected and therefore never be reported to an SRS.

A potential new way to detect the rare ADRs that cannot be identified by doctors or by current methods applied to SRS databases is to use The Health Improvement Network (THIN) database (www.thin-uk.com), an Electronic Healthcare Database (EHD) containing complete UK General Practice records for registered patients. The THIN database contains all medical events (such as illnesses, laboratory results, signs and symptoms or administrative events) that a doctor is informed of for a patient as well as their complete prescription histories. Therefore, any rare ADRs that are serious enough to be reported to a doctor are more likely to be detected at an earlier point in time by applying a suitable data mining method on the THIN database rather than mining the SRS databases.

Existing methods developed for the EHDs are often disproportionality based methods (methods that contrast how often the event of interest occurs after the specified drug relative to how often the event of interest occurs after any drug) similar to the SRS methods [17]–[19] or association rule mining methods [20], [21]. As EHDs do not contain links between drugs and suspected ADRs these are often inferred by investigating medical events that occur within some time period around a drug, but many of these medical events are linked to the cause of taking the drug and these "therapeutically related" medical events present a major issue with the majority of the existing methods. It has been demonstrated that these existing methods are currently not suitable for signaling rare ADRs [22]. However, it is the rare ADRs that are unlikely to be detected by mining SRS databases, due to underreporting, so developing an algorithm that can signal rare ADRs using the THIN database would be beneficial.

In this paper, we develop a novel computational metaanalysis framework that integrates the existing methods (MUTARA, HUNT, TPD, see Section III-A) and uses information obtained from the internet to efficiently and accurately identify rare ADRs that occur immediately or shortly after a drug is prescribed. The framework uses the dependency measures obtained from some of the existing electronic healthcare-based methods and novel values of interest as attributes for each medical event that occurs within a 30 day period after the drug of interest is prescribed for any patient. After the attributes are generated for each medical event we label some of the medical events by extracting information from the internet informing us of the indicator events and known ADRs for the drug of interest. The unlabeled medical events then have labels assigned by applying metric learning and semisupervised clustering. Finally, using the labels we develop a novel filter that removes medical events labeled as indicator events and then return the remaining medical events ordered by how often they occurred unpredictably within 30 days after the drug being investigated multiplied by weights based on their assigned labels.

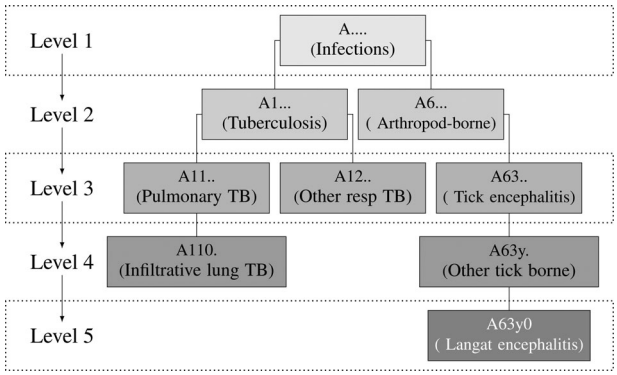

Fig. 1. Example of a branch within the Read Codes tree.

The continuation of this paper is as follows, Section II contains the background information on the THIN database and Section III describes the problem formulation. This is followed by the description of the novel methodology we developed that is able to identify rare ADRs in Section IV. Section V contains the results of comparing our novel algorithm with a selection of existing methods for the detection of known rare ADRs and the discussion of these results is contained in Section VI. The paper finishes with the conclusion and future work suggestions in Section VII.

II. THIN Database

The THIN database contains complete medical and prescription records for registered patients at participating general practices within the U.K. The medical information is recorded into the THIN database by Read Codes that correspond to illnesses, so each Read Code is paired with the illness description. Each Read Code is five elements long and the Read Codes have a tree structure. A level one Read Code has "." as its second element and corresponds to a very general description of an illness, for example "A. . . ." is a level one Read Code that corresponds to "Infections." A level two Read Code has "." for its third element but not for its second element, an example is "A1. . ." corresponding to "Tuberculosis." A level two Read Code is the child of the level one Read Code with a matching first element, so "A1. . ." is the child of "A. . . ." (or A. . . ." is the parent of "A1. . .") and a child Read Code corresponds to a more specific version of its parent's illness. Fig. 1 shows an example branch within the Read Code tree with the illness becoming more specific as the Read Code level increases.

Each drug is recorded into the THIN database by a drugcode that is paired to the generic name. The drugcode consists of nine numbers and does not have a structure we use but the drugcode does specify the way the drug is ingested and the dosage. Each entry also includes the date that a Read Code or drugcode is recorded but does not contain the time. In this paper, we used a subset of the THIN database containing records from 495 general practices. The subset contained approximately four million patients, more than 358 million prescription entries, and more than 233 million medical event entries.

Patients can register at a new practice at any point over their lifetime and it has been shown that statistical studies using the

TABLE I
EXAMPLES OF A MEDICAL TABLE AND A PRESCRIPTION TABLE CONTAINING THE PATIENTS' RECORDS

Drug Table			Medical Table		
patID	age	drugcode	patID	age	Read Code
jj3	10000	979596759	jj3	9999	A123.
jj3	20000	969686881	jj3	10000	F1...
jj3	20001	969686881	jj3	10000	C1...
jj3	20001	912314611	jj3	10001	A123.
aa2	15001	912314611	jj3	10013	D25..
aa2	15031	912314611	jj3	18020	C12..
aa2	15061	912314611	jj3	20001	C121.
aa2	25304	912314611	aa2	15001	B21..
bb8	10000	979596759	bb8	9078	A123.
			bb8	10000	F12..
			bb8	10000	A123.
			bb8	10010	D25..
			bb8	27002	C11..

The column patID is a unique reference corresponding to a patient and age is the patient's age in days when the entry was recorded.

THIN database will be biased if records from the first year after registration for a patient are included in the study [23]. The reason for this is that newly registered patients will need to inform their new doctor of any chronic illnesses they have, but these illnesses will be recorded on the day they inform the doctor of them rather than the actual day they were discovered. To prevent bias in this study we do not include the first year of a patient's medical records since registration. We also ignore the last 30 days of prescription records for each patient to reduce potential under reporting that may occur by including patients with less than 30 days of medical records after the first prescription of a drug.

III. PROBLEM FORMULATION

The focus of this paper is to develop a method for detecting rare ADRs that occur shortly after prescription, so we can restrict our attention to the medical events that occur around the time of the drug prescription. In this paper, we consider the natural numbers $\mathbb{N}$ to include 0 and use the interval $[a \dots b]$ to be the interval of natural numbers from a to b ($\{n \in \mathbb{N}\colon a \leq n \leq b\}$).

If we let M denote the set of Read Codes, D denote the set of drugcodes, and Ω denote the set of patients in the THIN database then the sequence of Read Codes or the sequence of drugcodes for a patient can be represented as two discrete functions. The discrete function representing a patient's Read Code sequence is a mapping $f_M\colon \mathbb{N} \times \Omega \to \mathcal{P}(M)$; $f_M(t,\omega) \to e_{\omega,t}$ where t is the age in days of the patient ω and $e_{\omega,t}$ is the set of Read Codes that are recorded into the THIN database for patient ω when they are t days old. Similarly, the discrete function representing a patient's drugcode sequence is $f_D\colon \mathbb{N} \times \Omega \to \mathcal{P}(D)$; $f_D(t,\omega) \to d_{\omega,t}$ where $d_{\omega,t}$ is the set of drugcodes recorded in THIN for the patient ω when they are t days old.

Using the example THIN database entries shown in Table I, we have

$$M = \{A123., B21., C1\dots, C11.., C12..,$$
$$C121., D25.., F1\dots, F12..\}$$
$$D = \{979\,596\,759, 969\,686\,881, 912\,314\,611\}$$
$$\Omega = \{jj3, aa2, bb8\}.$$

The set of Read Codes that patient $jj3$ has recorded when he is 9999, 10 000, and 10 002 days old are $f_M(9999, jj3) = \{A123.\}$, $f_M(10\,000, jj3) = \{F1\dots, C1\dots\}$, and $f_M(10\,002, jj3) = \{\} = \emptyset$, respectively. The reason $f_M(10\,002, jj3) = \{\} = \emptyset$ is that patient $jj3$ does not have a Read Code recorded when he is 10 002 days old. Similarly, the set of drugcodes that patient $jj3$ has recorded when he is 10 000 and 10 002 days old are $f_D(10\,000, jj3) = \{979\,596\,759\}$ and $f_D(10\,002, jj3) = \{\} = \emptyset$, respectively.

The set consisting of every age in days where patient ω has a drugcode recorded into the THIN database is

$$A_D(\omega) = \{t | f_D(t,\omega) \neq \emptyset\}.$$

The set of all drugs that are prescribed for patient ω is the finite union of the set of drugcodes recorded daily for the patient while they are active in the THIN database

$$\bigcup_{t \in A_D(\omega)} f_D(t,\omega)$$

so to determine the age that patient ω is first prescribed the drug of interest $d_* \in D$ we first find the set of ages that the patient was prescribed the drug

$$\alpha(\omega, d_*) = \{t \in A_D(\omega) | d_* \in f_D(t,\omega)\}$$

and then define a new function $\alpha_1\colon \Omega \times D \to \mathbb{N} \cup \{-1\}$:

$$\alpha_1(\omega, d_*) = \begin{cases} \min(\alpha(\omega, d_*)) & \text{if } d_* \in \bigcup\limits_{t \in A_D(\omega)} f_D(t,\omega) \\ -1 & \text{else} \end{cases}$$

that finds the minimum age that the patient is prescribed the drug or returns -1 if the patient has never been prescribed the drug. The set of patient's ages where the drug is prescribed for the first time in 13 months is determined by the function $\hat{\alpha}\colon \Omega \times D \to \mathcal{P}(\mathbb{N} \cup \{-1\})$,

$$\hat{\alpha}(\omega, d_*) = \{t \in \alpha(\omega, d_*) | (t - s^*) \geq 386,$$
$$s^* = \underset{\substack{s \in \alpha(\omega, d_*) \\ s < t}}{\operatorname{argmin}} (t - s)\} \cup \alpha_1(\omega, d_*).$$

If a patient is not prescribed the drug then the function returns -1 otherwise it returns the set of ages in days that the patient took the drug and had a minimum of 386 days between previously taking the drug.

For the example entries in Table I, the set of ages in days that patient $jj3$ is prescribed any drugs is $A_D(jj3) = \{10\,000, 20\,000, 20\,001\}$ and the set of ages in days that the patient is prescribed the specific drugs 979 596 759 and 969 686 881 is $\alpha(jj3, 979\,596\,759) = \{10\,000\}$ and $\alpha(jj3, 969\,686\,881) = \{20\,000, 20\,001\}$, respectively. The minimum age that patient $jj3$ prescribed 969 686 881 is $\alpha_1(jj3, 969\,686\,881) = 20\,000$ but as the patient $aa2$ is never prescribed the drug 969 686 881, $\alpha_1(aa2, 969\,686\,881) = -1$. The patient $aa2$ is prescribed the drug 912 314 611 four times ($\alpha(aa2, 912\,314\,611) = \{15\,001, 15\,031, 15\,061, 25\,304\}$) with the first prescription occurring at 15 001 days old ($\alpha_1(aa2, 912\,314\,611) = 15\,001$) but the drug was then repeated monthly two times and then the patient had a 10 243 day break before being prescribed the drug for the final time,

therefore the set of ages in days that patient $aa2$ is prescribed drug 912 314 611 for the first time in 13 months is $\hat{\alpha}(aa2, 912\,314\,611) = \{15\,001, 25\,304\}$ as the instances when the drug was repeated after 30 days are not included.

In the continuation of this paper, we will use $\hat{\alpha}(\omega, d_*)_k$ to refer to the kth time patient ω is prescribed drug d_* for the first time in 13 months, so $\hat{\alpha}(\omega, d_*)_1 < \hat{\alpha}(\omega, d_*)_2 < \cdots < \hat{\alpha}(\omega, d_*)_{n-1} < \hat{\alpha}(\omega, d_*)_n$, where n is the total number of times patient ω is prescribed drug d_* for the first time in 13 months ($n = |\hat{\alpha}(\omega, d_*)|$). Following on with our example, $\hat{\alpha}(aa2, 912\,314\,611)_1 = 15\,001$ and $\hat{\alpha}(aa2, 912\,314\,611)_2 = 25\,304$.

This then enables us to define a interval of interest around each time the patient is prescribed a drug for the first time in 13 months, for each $K \leq |\hat{\alpha}(\omega, d_*)|$,

$$T(\omega, d_*, t_1, t_2)_K = \begin{cases} [(\hat{\alpha}(\omega, d_*)_K + t_1) & \text{if } \hat{\alpha}(\omega, d_*)_K \neq -1 \\ \quad \ldots (\hat{\alpha}(\omega, d_*)_K + t_2)] & \\ \emptyset & \text{else} \end{cases} \quad (1)$$

(1) is defining a time interval centered around the Kth time drug d_* is prescribed for the first time in 13 months for patient ω determined by the integers t_1 and t_2. For example, if we wanted to investigate the 30 days after a prescription, we would use $t_1 = 1$ and $t_2 = 30$ and the time period around the first prescriptions in 13 months of drug 912 314 611 for patient $aa2$ would be $T(aa2, 912\,314\,611, 1, 30)_1 = [(\hat{\alpha}(aa2, 912\,314\,611)_1 + 1) \ldots (\hat{\alpha}(aa2, 912\,314\,611)_1 + 30)] = [(15\,001 + 1) \ldots (15\,001 + 30)] = [(15\,002) \ldots (15\,031)]$ and $T(aa2, 912\,314\,611, 1, 30)_2 = [(\hat{\alpha}(aa2, 912\,314\,611)_2 + 1) \ldots (\hat{\alpha}(aa2, 912\,314\,611)_2 + 30)] = [(25\,304+1) \ldots (25\,304+30)] = [(25\,305) \ldots (25\,334)]$. If the patient is never prescribed the drug then the time period of interest is the empty set, for example, $T(aa2, 979\,596\,759, 1, 30)_1 = \emptyset$.

As the function $f_M(t, \omega)$ returns a set of Read Codes that are recorded into the THIN database for patient ω when they are t days old, we can find the set of ages in days that ω has any Read Code recorded

$$A_M(\omega) = \{t | f_M(t, \omega) \neq \emptyset\}$$

and the finite union over each $t \in A_M(\omega)$,

$$\bigcup_{t \in A_M(\omega)} f_M(t, \omega)$$

is the set of all Read Codes that are recorded into the THIN database for the patient ω. It follows that $\bigcup_{t \in T(\omega, d_*, t_1, t_2)_K} f_M(t, \omega)$ is the set of all Read Codes that are recorded into the THIN database for the patient ω during the period of interest determined by t_1 and t_2 around the Kth prescription of drug d_*. The function $h\colon M \times D \times \Omega \times \mathbb{Z}^2 \to \{0, 1\}$,

$$h(e_i, d_*, \omega, t_1, t_2) = \mathbb{1}\left\{ e_i \in \bigcup_{t \in T(\omega, d_*, t_1, t_2)_1} f_M(t, \omega) \right\}$$

is one if the patient ω has the Read Code e_i recorded within the time period of interest around the first prescription of drug d_* and zero otherwise. To determine the number of times the drug d_* is prescribed to patient ω for the first time in 13 months and the Read Code e_i is recorded within the time period of interest we define another function $\hat{h}\colon M \times D \times \Omega \times \mathbb{Z}^2 \to \mathbb{N}$,

$$\hat{h}(e_i, d_*, \omega, t_1, t_2) = \sum_{K \leq |\hat{\alpha}(\omega, d_*)|} \mathbb{1}\left\{ e_i \in \bigcup_{t \in T(\omega, d_*, t_1, t_2)_K} f_M(t, \omega) \right\}.$$

The total number of patients in the THIN database that have Read Code e_i recorded within the time period of interest centered around the first prescription of drug d_* is then

$$\sum_{\omega \in \Omega} h(e_i, d_*, \omega, t_1, t_2)$$

and the total number of first times prescriptions of drug d_* in 13 months where the Read Code e_i is recorded within the time period centered around the prescription determined by t_1 and t_2 is

$$\sum_{\omega \in \Omega} \hat{h}(e_i, d_*, \omega, t_1, t_2).$$

A. Existing Algorithms

The temporal pattern discovery (TPD) [17] calculates the information component (IC) that looks at the disproportionality between how often a Read Code is recorded within some time period after the first prescription in 13 months of the drug of interest compared to how often it is recorded within the same time period after the first prescription in 13 months of any drug but also adds a bias lowering the IC value if the Read Code or drug is rare.

The function used to test if a set of medical events is empty is $\hat{H}\colon \mathcal{P}(M) \to \{0, 1\}$,

$$\hat{H}[B] = \begin{cases} 0 & \text{if } B = \emptyset \\ 1 & \text{else} \end{cases}$$

using the above function, we define the following values:

$$n_{d_* e_i}(t_1, t_2) = \sum_{\omega \in \Omega} \hat{h}(e_i, d_*, \omega, t_1, t_2)$$

$$n_{\bullet e_i}(t_1, t_2) = \sum_{d \in D} \sum_{\omega \in \Omega} \hat{h}(e_i, d, \omega, t_1, t_2)$$

$$n_{d_* \bullet}(t_1, t_2) = \sum_{\omega \in \Omega} \sum_{K=1}^{|\hat{\alpha}(\omega, d_*)|} \hat{H}\left[M \cap \left(\bigcup_{t \in T(\omega, d_*, t_1, t_2)_K} f_M(t, \omega) \right) \right]$$

$$n_{\bullet\bullet}(t_1, t_2) = \sum_{d \in D} \sum_{\omega \in \Omega} \sum_{K=1}^{|\hat{\alpha}(\omega, d)|} \hat{H}\left[M \cap \bigcup_{t \in T(\omega, d, t_1, t_2)_K} f_M(t, \omega) \right]$$

where $n_{d_* e_i}(0, 30)$ is the number of times that event e_i occurs in the month after a first prescription in 13 months of drug d_*, $n_{\cdot e_i}(0, 30)$ is the number of times that event e_i occurs in the month after a first prescription in 13 months of any drug, $n_{d_* \cdot}(0, 30)$ is the number of times a drug d_* is prescribed for the first time in 13 months and has any follow up in the month after the prescription, and $n_{\cdot\cdot}(0, 30)$ is the number of times any drug is prescribed for the first time in 13 months and has any follow up in the month after the prescription.

The expected number of first time prescriptions of drug d_* in 13 months that have the Read Code e_i occur within the time pe-

riod defined by t_1 and t_2 if the Read Code occurs independently of the drug is

$$E_{d_*e_i}(t_1,t_2)=\frac{n_{d_*\bullet}(t_1,t_2)n_{\bullet e_i}(t_1,t_2)}{n_{\bullet\bullet}(t_1,t_2)}$$

and the $IC(t_1,t_2,e_i,d_*)$ value calculates how many first prescription in 13 months of drug d_* have the Read Code e_i occurring between t_1 and t_2 days after divided by the expected value if the Read Code and drug occur independently

$$IC(t_1,t_2,e_i,d_*)=\log_2\left(\frac{n_{d_*e_i}(t_1,t_2)+1/2}{E_{d_*e_i}(t_1,t_2)+1/2}\right).$$

The value of a half added to both the numerator and denominator in the $IC(t_1,t_2,e_i,d_*)$ value calculation creates a bias that causes the $IC(t_1,t_2,e_i,d_*)$ value to tend to zero when the Read Code or drug occurs rarely.

The IC_Δ is a measure that compares the $IC(0,30,e_i,d_*)$ corresponding to the time period of a month after the first prescription in 13 months compared to the $IC(-810,-630,e_i,d_*)$ corresponding to a time period between 27 and 21 months prior to the prescription

$$IC_\Delta(e_i,d_*)=\log_2\left(\frac{n_{d_*e_i}(0,30)+1/2}{E^*_{d_*e_i}(0,30)+1/2}\right) \tag{2}$$

where

$$E^*_{d_*,e_i}(0,30)=\frac{n_{d_*e_i}(-810,-630)E_{d_*e_i}(0,30)}{E_{d_*e_i}(-810,-630)}.$$

When investigating possible ADRs for drug d_* the authors use a filter to remove any Read Codes e_i if the $IC(-30,-1,e_i,d_*)>IC(0,30,e_i,d_*)$ or $IC(0,0,e_i,d_*)>IC(0,30,e_i,d_*)$, these correspond to the IC value the month prior to the prescription being higher than the month after prescription, or the IC value on the day of prescription (DOP) being higher than a month after. In this paper, we will use the $IC_\Delta(e_i,d_*)$ as described above as an attribute for each possible drug and Read Code combination and use the Heaviside step function $H\colon\mathbb{R}\to\{0,1\}$,

$$H[n]=\begin{cases}0 & \text{if } n\le 0\\ 1 & \text{if } n>0\end{cases}$$

to define the filter functions

$$\begin{aligned}\zeta_1(e_i,d_*)&=H[IC(-30,-1,e_i,d_*)-IC(0,30,e_i,d_*)]\\ \zeta_2(e_i,d_*)&=H[IC(0,0,e_i,d_*)-IC(0,30,e_i,d_*)]\end{aligned} \tag{3}$$

as two addition binary attributes.

The algorithm Mining Unexpected Temporary Association Rules given the Antecedent (MUTARA) [20] applies a case control approach that estimates the background rate that the Read Code is recorded into the THIN database by finding out how many patients who have not been prescribed the drug of interest have the Read Code recorded during a random time interval. MUTARA aims to find Unexpected Temporary Association Rules (UTARs) by investigating how many patients have a specific Read Code unexpectedly recorded within a period of interest centered on the first prescriptions of the drug being studied. The algorithm Highlighting UTARs Negating TARs (HUNT) [21] was developed by the same authors as MUTARA and applies a similar method but is less prone to ranking therapeutic failure Read Codes (Read Codes linked to the cause of taking the drug) above Read Codes corresponding to ADRs.

MUTARA and HUNT both investigate the Read Codes that occur during the month after the drug being studied is first prescribed to a patient or the union of the month after the first and second prescriptions of the drug if it is repeated within a month of the first prescription. Previously, we defined $\alpha(\omega,d_*)$ to be the ages in days that patient ω is prescribed the drug d_* and $\alpha_1(\omega,d_*))$ to be the age in days that patient ω is first prescribed the drug d_*, so the set $\alpha(\omega,d_*)\setminus\alpha_1(\omega,d_*)$ contains the ages that patient ω is prescribed drug d_* except the age when they are first prescribed. Using this, we define a new function that returns the patient's age in days for the second time a drug is prescribed for a patient if this is within a month of the first prescription or the patient's age in days for the first time the drug is prescribed if the second prescription is not within a month of the first

$$\alpha_2(\omega,d_*)=\begin{cases}\overset{*}{t}=\underset{t\in\alpha(\omega,d_*)\setminus\alpha_1(\omega,d_*)}{\arg\min}(|t-\alpha_1(\omega,d_*)|) & \text{if } \overset{*}{t}<30\\ \alpha_1(\omega,d_*) & \text{else}\end{cases}$$

MUTARA uses patients that have not been prescribed the drug (so $\alpha_1(\omega,d_*)=-1$) to estimate the background rate that the Read Code is prescribed into the THIN database. For each patient that has never been prescribed the drug d_*, a random time interval is chosen within the age that the patient first has any Read Code recorded and the age that a Read Code is last recorded, $\min(A_M(\omega))$ and $\max(A_M(\omega))$, respectively, where we defined $A_M(\omega)$ previously to be the set of ages that patient ω has a Read Code recorded into the THIN database. If a patient has never had a Read Code recorded into the THIN database ($A_M(\omega)=\emptyset$) or only had Read Codes recorded for the short period of time ($\max(A_M(\omega))-\min(A_M(\omega))<|t_2-t_1|$) then the patient is not used in this study as they are not an active patient and may bias results. The time interval of length $|t_2-t_1|$ is chosen uniformly within $[\min(A_M(\omega))\ldots\max(A_M(\omega))]$.

Putting this all together, the time interval of interest around the first prescription of drug d_* for patients prescribed the drug or the time interval chosen at random for patients who have never been prescribed the drug used by MUTARA is

$$T_M(\omega,d_*,t_1,t_2)=\begin{cases}[(\alpha_1(\omega,d_*)+t_1) & \text{if } \alpha_1(\omega,d_*)\neq -1\\ \quad\ldots(\alpha_2(\omega,d_*)+t_2)] & \\ [(\min(A_M(\omega))+r) & \text{else}\\ \quad\ldots(\min(A_M(\omega))+r+|t_2-t_1|)] & \end{cases}$$

where

$$r\sim U(0,\max(A_M(\omega))-\min(A_M(\omega))-|t_2-t_1|).$$

For patients who are prescribed the drug d_* a filter is applied to ignore any "expected" Read Codes that are recorded within the time interval of interest after the first prescription of the drug. A Read Code is "expected" for the patient during the time interval of interest after the drug is prescribed if the patient also had the Read Code recorded within $t_3\in\mathbb{N}$ days prior to drug being prescribed. We define this time interval prior to the

prescription as

$$T_{\text{filt}}(\omega, d_*, t_3) = \begin{cases} [(\alpha_1(\omega, d_*) - t_3) \dots (\alpha_1(\omega, d_*) - 1)] & \text{if } \alpha_1(\omega, d_*) \neq -1 \\ \emptyset & \text{else.} \end{cases}$$

We then determine if patient ω has Read Code e_i recorded "unexpectedly" during the time interval of interest after drug d_* by

$$h^*(e_i, d_*, \omega, t_1, t_2, t_3) = \mathbb{1}\left\{ e_i \in \underset{t \in T_M(\omega, d_*, t_1, t_2)}{\cup f_M(t, \omega)} \right\} \times \left(1 - \mathbb{1}\left\{ e_i \in \underset{t \in T_{\text{filt}}(\omega, d_*, t_3)}{\cup f_M(t, \omega)} \right\}\right).$$

The Heaviside step function applied to $\alpha_1(\omega, d_*)$ returns 1 if patient ω has been prescribed the drug d_*, as in that case $\alpha_1(\omega, d_*) > 0$ or returns 0 if patient ω has never been prescribed the drug d_*. MUTARA then calculates the unexpected-leverage as

$$uL_{e_i}^{d_*} = \frac{\sum_{\omega \in \Omega} [h^*(e_i, d_*, \omega, t_1, t_2, t_3) \times H[\alpha_1(\omega, d_*)]]}{|\Omega|} - \left(\frac{\sum_{\omega \in \Omega} H[\alpha_1(\omega, d_*)]}{|\Omega|}\right)\left(\frac{\sum_{\omega \in \Omega} h^*(e_i, d_*, \omega, t_1, t_2, t_3)}{|\Omega|}\right)$$

where the fraction of patients in the database who have the drug multiplied by the fraction of patients who have the Read Code recorded during their interval of interest is subtracted from the fraction of patients in the database who have the Read Code recorded "unexpectedly" within a month of the first time the drug is prescribed (or up to a month after the second prescription if the drug is repeated within a month).

HUNT calculates a similar value to the unexpected-leverage known as the leverage that does not include a filter to remove "expected" Read Codes based on a patient's history

$$h^{**}(e_i, d_*, \omega, t_1, t_2) = \mathbb{1}\left\{ e_i \in \underset{t \in T_M(\omega, d_*, t_1, t_2)}{\cup f_M(t, \omega)} \right\}$$

and

$$L_{e_i}^{d_*} = \frac{\sum_{\omega \in \Omega} [h^{**}(e_i, d_*, \omega, t_1, t_2) \times H[\alpha_1(\omega, d_*)]]}{|\Omega|} - \left(\frac{\sum_{\omega \in \Omega} H[\alpha_1(\omega, d_*)]}{|\Omega|}\right)\left(\frac{\sum_{\omega \in \Omega} h^{**}(e_i, d_*, \omega, t_1, t_2)}{|\Omega|}\right).$$

HUNT then ranks the Read Codes in descending order of the leverage to attain the leverage rank and in descending order of the unexpected-leverage to attain the unexpected-leverage rank and then calculates the rank ratio between the leverage rank and unexpected-leverage rank. Finally, HUNT returns the Read codes in descending order of this rank ratio.

B. Proposed Algorithm Attributes

In the previous sections, we have defined the problem and described three existing algorithms. In this section, we develop novel attributes that may help a semisupervised algorithm separate Read Codes corresponding to ADRs from noise or indications. The first novel attribute of interest for detecting ADRs is how many patients have the Read Code recorded a month prior to the first prescription compared with how many patients have it recorded a month after

$$ABratio_{30}(e_i, d_*) = \frac{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, 1, 30)}{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, -30, -1)} \quad (4)$$

a similar new attribute is to consider the ratio between the number of patients who have the Read Code recorded a year after compared to a year before

$$ABratio_{365}(e_i, d_*) = \frac{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, 1, 365)}{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, -365, -1)}. \quad (5)$$

As the THIN database does not record the time that a Read Code or drugcode is recorded, a Read Code that are recorded on the same day as the drugcode could correspond to a serious ADR that occurs immediately or an indicator, this is why the DOP is not included in the above $ABratios$. To incorporate data from the DOP we define another attribute by

$$DOP(e_i, d_*) = \frac{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, 0, 0)}{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, -365, -1)}. \quad (6)$$

In previous methods, there has been a patient level filter that removes medical events from a patient's sequence that occur 30 days after the drug if the patient also had the medical event shortly prior to the drug [20]. The justification for this is that an illness is unlikely to be an ADR if it occurred shortly before the drug and then repeatedly occurred after the drug for the same patient. This inspires the third attribute that is based on the number of patients who have the Read Code recorded in the month after and also have it recorded during the month prior to the first prescription

$$ex(e_i, d_*) = \frac{\sum_{\omega \in \Omega} [h(e_i, d_*, \omega, 1, 30) h(e_i, d_*, \omega, -30, -1)]}{\sum_{\omega \in \Omega} h(e_i, d_*, \omega, 1, 30)}. \quad (7)$$

The final attributes of interest make use of the Read Code tree structure. When a patient first has an illness it is likely that not much detail is known, so a low level Read Code will probably be recorded into the THIN database, over time more information may be discovered about a patient's illness possibly due to laboratory results and this may then result in a more specified higher level Read Code being entered (a child of the original less detailed Read Code). As a consequence, it is common for higher level Read Codes related to the cause of taking the drug to only be recorded after the drug, so to help distinguish between these and ADRs we calculate the Read Code after and before ratio when only considering the first two or three elements of a Read Code. For example, considering the first three elements of the Read Code, the Read Code $A11ab$ becomes $A11$ and the AB_{lev3} is the number of patients that have a Read Code starting with $A11$ within the month after the first prescription of the drug divided by the number of patients that have a Read Code starting with $A11$ within the month prior to the first prescription of the drug. Similarly, the AB_{lev2} for a Read Code is the number of patients who have a Read Code with the same first two elements within a month after the first prescription divided by the number of patients who have a Read Code with the same first two elements within a month prior to the prescription.

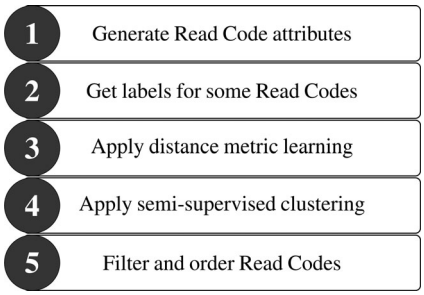

Fig. 2. Five steps applied in the DRESS algorithm.

To calculate these attributes, we first define an equivalence relationship, $e_i \overset{n}{\sim} e_j$ if the first n elements of the Read Codes of e_i and e_j are the same. The number of patients who have Read Codes equivalent to e_i recorded a month after the first prescription of d_* is

$$\sum_{\omega\in\Omega} H\left[\sum_{e\overset{n}{\sim}e_i} h(e,d_*,\omega,1,30)\right]$$

similarly, the number of patients who have Read Codes equivalent to e_i recorded during the month prior to the prescription is

$$\sum_{\omega\in\Omega} H\left[\sum_{e\overset{n}{\sim}e_i} h(e,d_*,\omega,-30,-1)\right].$$

It follows that the AB_{lev3} and AB_{lev2} are calculated by

$$AB_{lev3}(e_i,d_*) = \frac{\sum_{\omega\in\Omega} H[\sum_{e\overset{3}{\sim}e_i} h(e,d_*,\omega,1,30)]}{\sum_{\omega\in\Omega} H[\sum_{e\overset{3}{\sim}e_i} h(e,d_*,\omega,-30,-1)]} \quad (8)$$

$$AB_{lev2}(e_i,d_*) = \frac{\sum_{\omega\in\Omega} H[\sum_{e\overset{2}{\sim}e_i} h(e,d_*,\omega,1,30)]}{\sum_{\omega\in\Omega} H[\sum_{e\overset{2}{\sim}e_i} h(e,d_*,\omega,-30,-1)]}. \quad (9)$$

IV. DRESS Algorithm

The detecting rare events semisupervised (DRESS) algorithm comprises of five steps, see Fig. 2. The DRESS algorithm requires the user to input the drug of interest (d_*) and returns a ranked list of Read Codes in descending order of how likely they are to be ADRs. Tentative ADR signals can then be determined by the DRESS algorithm by considering the top k ranked Read codes to be signaled as ADRs. The value of k will determine the filtering threshold and in this paper we consider the top 100 ranked Read Codes to be signaled by the algorithms.

A. Step 1

The first step of the DRESS algorithm is the generation of attributes for any Read Code that could be an ADR. This is accomplished by initially finding the set of all the Read Codes that are recorded within a month of the first prescription of d_* for any patient, we denote this set of Read Codes by G. Then, for every Read Code $e_i \in G$ the DRESS algorithm generates the novel attributes $ABratio_{30}(e_i,d_*), ABratio_{365}(e_i,d_*), DOP(e_i,d_*), ex(e_i,d_*), AB_{lev3}(e_i,d_*), AB_{lev2}(e_i,d_*)$, see (4)–(9), and the existing TPD values the $IC_\Delta(e_i,d_*), \zeta_1(e_i,d_*)$ and $\zeta_2(e_i,d_*)$, see (2)–(3). The data point for each $e_i \in G$ is denoted by $\mathbf{x_i} \in \mathbb{R}^9$ and is a vector corresponding to the attribute values. The set of data points is denoted by $X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, where the cardinality of the set G is n ($|G| = n$).

B. Step 2

As we are applying a semisupervised approach, we need labels for some of the Read Codes. We have decided to have three different labels for each Read Code, one label representing Read Codes that are ADRs (known ADR), another label representing Read Codes that cause the drug to be taken by the patients (indicator), and the final label representing Read Codes that are not linked to the drug but just occur by chance (noise). The reason for choosing three labels is because there is information available to enable us to determine the labels for a sufficient number of Read Codes when using three labels.

We determined the Read Codes that are labeled as noise by using the hierarchal structure of the Read Code tree. We determined branches that are not related to immediately occurring ADRs by manually investigating the Read Code tree and found the set of irrelevant Read Codes M_{irrel}. Examples of Read Codes in M_{irrel} that cannot correspond to immediately occurring ADRs are those related to cancer, occupations or family history. Fig. 3 illustrates the Read Code tree and the Read Code branches considered to be noise are shaded in, whereas the Read Code branches that are possible ADRs are unshaded. Using the set M_{irrel}, the DRESS algorithm labels any Read Code $e_i \in G \cap M_{\text{irrel}}$ as noise.

To determine the labels of some of the Read Codes corresponding to known ADRs or indicators, we mined data from the internet. The Read Codes labeled as indicators are found by first extracting the strings listed as indicators on the NetDoctor website [24], then finding the set of Read Codes with descriptions containing any one of these strings and finally validating these are indicators by ignoring indicator Read Codes that do not have an $ABratio_{30} < 1$. The Read Codes labeled as known ADRs are found similarly, by first extracting the strings listed as side effects on the NetDoctor website, then finding the set of Read Codes with descriptions containing any one of these strings and validating these by ignoring any Read Codes that do not have an $ABratio_{30} \geq 1.5$. In addition to the Read Codes corresponding to the NetDoctor listed side effects, Read Codes with a description containing the drug name and the term "adverse" are also labeled as known ADRs.

C. Step 3

After labeling some of the Read Codes we then applied the metric learning algorithm detailed in [25]. Letting S denote the set of all index pairs for Read Codes with the same label (e.g., if Read Codes corresponding to the data points $\mathbf{x_1}$ and $\mathbf{x_2}$ are both labeled as noise then $(1,2) \in S$), D denote the set of all index pairs for Read Codes with a different label (e.g., if the Read code corresponding to data point $\mathbf{x_1}$ is labeled as noise but the Read code corresponding to the data point $\mathbf{x_5}$ is labeled as an indicator then $(1,5) \in D$) and the inner product

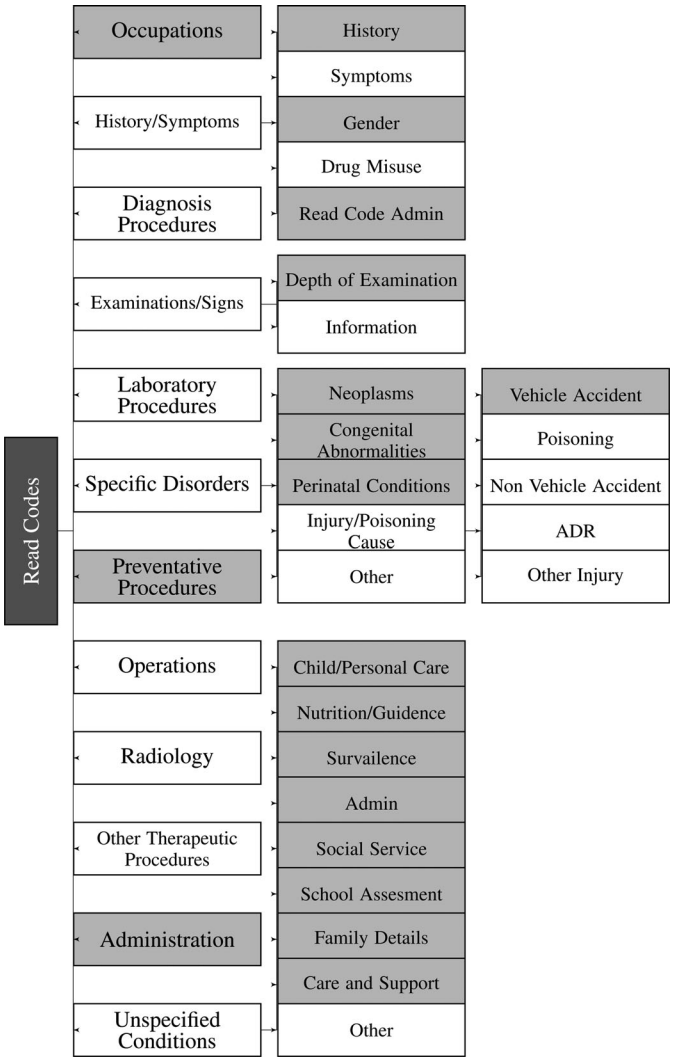

Fig. 3. Read Codes tree.

of two $d \times n$ real valued matrices, $X, Y \in \mathbb{R}^{d\times n}$, is denoted by $\langle X, Y\rangle := \mathbf{Tr}(\mathbf{X^T Y})$, where **Tr**(A) means the trace of the matrix A. The cone of positive semidefinite matrices is denoted by S_+^d.

Given a pair of Read Code data points $\mathbf{x_i}$ and $\mathbf{x_j}$ generated in step 1, we calculate 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 $\tilde{X}_\tau = X_S^{-1/2} X_\tau X_S^{-1/2}$. The authors calculated that $\nabla f_\mu(S_t^\mu) = \frac{\sum_{\tau\in D} e^{-\langle \tilde{X}\tau, S\rangle/\mu} \tilde{X}_\tau}{\sum_{\tau\in D} e^{-\langle \tilde{X}\tau, S\rangle/\mu}}$.

When applied in the DRESS algorithm, the distance metric algorithm described in algorithm 1 finds a mapping from the space consisting of the attribute values found in step 1 to a different space that minimizes the distance between two Read Codes with the same label and maximizes the distance between two Read Codes with different labels. The mapping is, $f\colon \mathbb{R}^9 \to \mathbb{R}^9$; $f(\mathbf{x_i}) = \mathbf{x_i}^T S_t^\mu$, where S_t^μ is the 9×9 learned distance metric matrix.

Algorithm 1 The metric learning algorithm [25]

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 **Tr**(S_1^μ)= 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

D. Step 4

The constrained K-means algorithm [26] is applied to the Read Codes on their corresponding transformed data points determined by the distance metric learning algorithm above. The constrained K-means algorithm is a semisupervised algorithm that fixes the class of the labeled Read Codes and uses these labeled Read Codes to calculate the initial cluster centers then iteratively assigns the nonfixed Read Codes into the cluster with the closest mean, with the means iteratively being recalculated until convergence.

When the DRESS algorithm is implemented, the set of data points input into the constrained K-means algorithm is the set $\{f(\mathbf{x_1}), f(\mathbf{x_2}), \ldots, f(\mathbf{x_n})\}$, the value of K input is 3 and the initial seeds are $S_1 = \{\mathbf{x_i}\colon \mathbf{x_i}$ is labeled as a known ADR$\}$, $S_2 = \{\mathbf{x_i}\colon \mathbf{x_i}$ is labeled as an indicator$\}$ and $S_3 = \{\mathbf{x_i}\colon \mathbf{x_i}$ is labeled as noise$\}$.

Algorithm 2 The Constrained K-means algorithm [26]

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*;

Read Codes in the same cluster as the Read Codes that were originally labeled as known ADRs are referred to as being in the ADR cluster, Read Codes in the same cluster as the Read

TABLE II
READ CODE RANKS RETURNED BY THE DIFFERENT METHODS

Drug	ADR (Year withdrawn)	Incidence	DRESS	TPD	HUNT	MUTARA
Rofecoxib	Inferior myocardial infarction (2004)	0.79-1.24% [27]	83	302	196	1212
Rofecoxib	Heart failure (2004)	0.79-1.24% [27]	121	805	504	562
Rimonabant	Depressed mood (2008)	0.6% [28]	16	32	15	68
Rimonabant	Depressive disorder NEC (2008)	3 % [29]	23	180	295	1157
Enalapril	Acute pancreatitis	rare [30], [31]	53	700	416	2081
Naproxen	Acute renal failure	rare [32]	20	159	138	1313
Naproxen	Jaundice	rare [33]	52	774	3394	4190
Naproxen	Hepatitis unspecified	rare [33]	241	35	852	3502
Naproxen	Suicide and selfinflicted injury	rare [34]	104	223	242	1808
Nifedipine	Weight increasing	$\leq 1\%$ [35]	362	29	1517	1492
Nifedipine	Feels hot/feverish	$\leq 2\%$ [35]	314	428	1972	7480
Ciprofloxacin	Nontraumatic rupture of Achilles tendon	rare [36]	13	42	1273	5506
Levofloxacin	Rupture Achilles tendon	rare [36]	76	296	-	-
Levofloxacin	Nontraumatic rupture of Achilles tendon	rare [36]	298	279	320	1189
Rifampicin	Thrombocytopenia NOS	rare [37]	75	418	265	1003
Metformin	Lactic acidosis	0.005 % [38]	-	-	-	-
Metformin	Acidosis	0.005 % [38]	449	813	466	3223

If the algorithm did not return a rank then this is represented by '-'.

Codes that were originally labeled as indicators are referred to as being in the indicator cluster and Read Codes in the same cluster as the Read Codes that were originally labeled as noise are referred to as being in the noise cluster.

E. Step 5

The last step involved applying two additional filters and then used the Read Code attributes and clustering to order the Read Codes by how likely they are to be ADRs. The first filter removed all the Read Codes (e_is) that were in the indicator cluster or where $(1 - ex(e_i, d_*)) \times ABratio_{30}(e_i, d_*) < 1$. The second filter is a filter that we have developed for postprocessing with any algorithm that detects ADRs by mining the THIN database. This filter removes all the Read Codes that are irrelevant for ADR detection such as Read Codes corresponding to administrative events. Finally, Read Codes were ordered in descending order of $(1 - ex(e_i, d_*)) \times ABratio_{30}(e_i, d_*) \times \frac{1}{\beta}$, where $\beta = 1$ for Read Codes in the ADR cluster and $\beta = 3$ for Read Codes in the noise cluster.

If DRESS returned the ranked list of all the Read Codes then this would consist of the labeled and unlabeled ones. The rank of the labeled Read Codes is irrelevant as we already know whether they represent ADRs. As a consequence, we decided to ignore the labeled Read Codes and only returned the unlabeled Read Codes in the ranked list. In summary, the DRESS algorithm uses the semisupervised clustering to filter Read Codes that have attributes that make them unlikely to be ADRs and then orders the remaining Read Codes by how often they occurred unexpectedly after the first prescription of the drug being investigated but also adds a weight so that Read Codes that have attributes similar to known ADRs are ranked higher.

V. RESULTS

We applied the DRESS algorithm to a range of drugs, some of which have been withdrawn from the market due to serious ADRs. Table II shows the rank that each data mining algorithm assigns for a known rare and serious drug and ADR pair. The table also states the year the drug was withdrawn and approximately how common the ADR is. In some cases the rare and serious ADR being investigated was listed as an ADR on the NetDoctor website and to prevent any bias in the results for the DRESS algorithm, any labels for the Read Codes corresponding to the ADR being investigated were removed at the end of step 2, prior to the semisupervised steps. So the Read Codes corresponding to the ADR being investigated were always unlabeled in the DRESS algorithm. As DRESS only ranks the unlabeled Read Codes, to enable a fair comparison and prevent bias, we also filtered the labeled Read Codes from the ranked lists returned by the existing methods.

The DRESS algorithm had an average rank over the drug and ADR pairs of 143.75 and had the highest rank compared to all the other algorithms for 12 of the 16 drug and ADR pairs investigated. It was able to return a rank under 100 for 56.25% of the ADRs and all the ADRs had a rank below 500. In comparison the other methods had an average rank of 344.69, 791, and 2385.73 for TPD, HUNT, and MUTARA, respectively. The TPD only returned a rank under 100 for 25% of the ADRs and only 6.25% of ADRs ranked by MUTARA and HUNT had a rank under 100.

VI. DISCUSSION

The existing methods for detecting ADRs using EHDs can be considered to generate tentative ADR signals for the top k ranked medical events in their returned list. This is effectively filtering out all the medical events that are unlikely to be ADRs or the medical events without sufficient evidence (number of patients experiencing the event after the drug) of being an ADR. The medical events with a rank greater than k are ignored and the medical events with a rank less than k (the signaled medical events) are investigated further to confirm if they are true ADRs. Therefore, for an unknown ADR to be detected by these existing methods it needs to be signaled by being ranked in the top k medical events returned and the closer that the rank is to the value 1, the more likely it will be investigated further, even when low values of k are used. The results of this paper show that the existing methods are not able to rank the known rare and

serious ADRs highly, so they would be unlikely to signal these for further investigation (or if they did get signaled, they would be surrounded by superfluous false signals), but the DRESS algorithm was able to rank more than 50% of these rare ADRs within the top 100 and would most likely signal these for further investigation. This implies that the DRESS algorithm is more suitable than existing methods for detecting rare ADRs and has the potential to discover many unknown rare ADRs.

The DRESS algorithm was unable to rank "Naproxen and Hepatitis" and "Nifedipine and weight increase" higher than the TPD, one reason for this is that the DRESS algorithm does not perform as well for medical events that have a high background rate as medical events that are common will have a greater number of patients experiencing the medical event in the month before the prescription and if the ADR is rare only one or two patients extra will have the medical event in the month after the prescription, so the $ABratio$ will be close to one, but on the other hand, the TPD performs better for medical events with a high background rate as the bias in the IC calculation will have less impact. It might be better to have two ADR clusters, one for the medical events that occur at a low background rate and another for the medical events that occur at a high background rate, but this may cause issue if there is not a sufficient number of Read Codes corresponding to known side effects, as if there are only a few Read Codes for the known side effects there may only be one to two labeled Read Codes in each cluster and this will be deleterious for the semisupervised methods.

It may be argued that the DRESS algorithm can only be applied after many ADRs are known and this may prevent it efficiently detecting rare ADRs, but rare ADRs that occur less than 1 in 1000 patients generally need three of more ADR cases before there is satisfactory evidence to confirm the ADR and this requires thousands of patients having the drug. DRESS can be implemented after a few hundred patients have taken a drug as the obvious side effects will be known, so the constraint of requiring known ADRs will not reduce the efficiency of DRESS. It is worth noting that the current implementation of DRESS will not be as effective if a drug is generally safe and does not have many side effects, but this is not common and DRESS could be modified to use the positive effects of the drug as labeled medical events as the positive effects and side effects have similar attribute values as both should increase after the drug is taken.

The DRESS algorithm still highly ranks medical events related to the cause of the drug and removing these medical events would greatly improve the ability to detect rare ADRs with DRESS. The ex attribute has limitations as it only indicates if the patient had a repeat of a medical event and does not make use of the medical event relations within the THIN database to determine if a medical event is expected based on related medical events. For example, if a patient has "a cold" then they are likely to experience "a cough" if the illness progresses, but the ex attribute only says "a cough" is expected if the patient has had "a cough" shortly before and not if a predecessor medical event has occurred before. To reduce the rank of medical events related to the cause of the drug a new attribute that uses sequential patterns to determine the expectedness of each medical event that occurs after the description could be used.

One possible bias in this study is that we cannot be 100% confident that the ADRs of a drug that are listed in medical sources are true due to the difficulty in determining causality. In this paper, we have assumed that medical events listed as ADRs on NetDoctor, which occur 1.5 times more often in the month after the prescription than the month before in the THIN database, are true ADRs. This assumption is not completely accurate as confounding effects may make some non-ADR medical events appear like ADRs. However, the more sources that confirm an ADR signal the more likely the signal is to be true, so by adding in the additional validation of only using medical events that have an $ABratio_{30}$ greater than 1.5 (are effectively signaled in the THIN database) we are reducing the likelihood that the labeled ADRs are incorrect. A possible way to further reduce the chance of using incorrect labels would be to incorporate additional validation steps, one such example could be to use peoples search history logs to see if there is evidence of people suspecting the ADRs [39].

VII. Conclusion

In this paper, we have described a novel methodology to detect rare ADRs that incorporates some existing methods (TPD, MUTARA, and HUNT), information retrieval from the web, metric learning, and semisupervised clustering. The results suggest this methodology is able to detect rare and serious ADRs for the range of drugs chosen in the investigation and has the potential to help detect many currently unknown ADRs. The method is not able to remove all medical events related to the cause of taking the drug and future work should aim to prevent generating signals for these medical events by adding an additional attribute to the clustering that determines the expectedness of each medical event based on sequential patterns that can be mined from the whole database. DRESS, as presented here, limits its risk windows to no more than 30 days beyond the original exposure. However, ADRs may have a much longer time between first exposure and ADR occurrence, and the risk may increase with cumulative exposure. The reason that the 30 day risk period was chosen in this paper was to enable a fair comparison with other published methods, which use a similar time window [40]. It would be straightforward to adapt DRESS for a more general risk window, but this is outside the scope of this paper and an area of future research that has had little focus to date.

References

[1] S. Gautier, H. Bachelet, R. Bordet, and J. Caron, "The cost of adverse drug reactions," *Expert Opin. Pharmacother.*, vol. 4, no. 3, pp. 319–326, 2003.

[2] K. J. Patel, M. S. Kedia, D. Bajpai, S. S. Mehta, N. A. Kshirsagar, and N. J. Gogtay, "Evaluation of the prevalence and economic burden of adverse drug reactions presenting to the medical emergency department of a tertiary referral centre: A prospective study," *BioMed. Central Pharmacol. Toxicol.*, vol. 7, no. 1, pp. 8–12, 2007.

[3] 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 Sci. One*, vol. 4, no. 2, p. e4439, 2009, doi: 10.1371/journal.pone.0004439.

[4] K. A. Hartholt, N. van der Velde, C. W. Looman, M. J. Panneman, E. F. Van Beeck, P. Patka, and T. J. van der Cammen, "Adverse drug reactions related hospital admissions in persons aged 60 years and over,

the Netherlands, 1981–2007: less rapid increase, different drugs," *Public Library Sci. One*, vol. 5, no. 11, 2010, doi: 10.1371/journal.pone.0013977.
[5] G. Shepherd, P. Mohorn, K. Yacoub, and D. W. May, "Adverse drug reaction deaths reported in united states vital statistics, 1999–2006," *Ann. Pharmacother.*, vol. 46, no. 2, pp. 169–175, 2012.
[6] T. M. Betteridge, C. M. Frampton, and D. L. Jardine, "Polypharmacy—We make it worse! a cross-sectional study from an acute admissions unit," *Internal Med. J.*, vol. 42, no. 2, pp. 208–211, 2012.
[7] 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," *Brit. Med. J.*, vol. 329, no. 7456, pp. 15–19, 2004.
[8] F. R. Varallo, M. F. R. Lima, J. C. F. Galduróz, and P. C. Mastroianni, "Adverse drug reaction as cause of hospital admission of elderly people: A pilot study," *Latin Amer. J. Pharmacy*, vol. 30, no. 2, pp. 347–353, 2011.
[9] J. A. Berlin, S. C. Glasser, and S. S. Ellenberg, "Adverse event detection in drug development: Recommendations and obligations beyond phase 3," *Amer. J. Public Health*, vol. 98, no. 8, pp. 1366–1371, 2008.
[10] B. Seruga, L. Sterling, L. Wang, and I. F. Tannock, "Reporting of serious adverse drug reactions of targeted anticancer agents in pivotal phase iii clinical trials," *J. Clin. Oncol.*, vol. 29, no. 2, pp. 174–185, 2011.
[11] L. Bergman, M. L. Beelen, M. P. Gallee, H. Hollema, J. Benraadt, and F. E. van Leeuwen, "Risk and prognosis of endometrial cancer after tamoxifen for breast cancer," *Lancet*, vol. 356, no. 9233, pp. 881–887, 2000.
[12] L. A. Ladewski, S. M. Belknap, J. R. Nebeker, O. Sartor, E. A. Lyons, T. C. Kuzel, M. S. Tallman, D. W. Raisch, A. R. Auerbach, G. T. Schumock *et al.*, "Dissemination of information on potentially fatal adverse drug reactions for cancer drugs from 2000 to 2002: First results from the research on adverse drug events and reports project," *J. Clin. Oncol.*, vol. 21, no. 20, pp. 3859–3866, 2003.
[13] S. J. W. Evans, P. C. Waller, and S. Davis, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports," *Pharmacoepidemiol. Drug Safety*, vol. 10, no. 6, pp. 483–486, 2001.
[14] A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, "A Bayesian neural network method for adverse drug reaction signal generation," *Eur. J. Clin. Pharmacol.*, vol. 54, no. 4, pp. 315–321, 1998.
[15] W. DuMouchel, "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system," *Amer. Statist.*, vol. 53, no. 3, pp. 177–190, 1999.
[16] L. Hazell and S. A. W. Shakir, "Under-reporting of adverse drug reactions," *Drug Safety*, vol. 29, no. 5, pp. 385–396, 2006.
[17] G. N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards, "Temporal pattern discovery in longitudinal electronic patient records," *Data Mining Knowl. Discov.*, vol. 20, no. 3, pp. 361–387, 2010.
[18] M. J. Schuemie, "Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD," *Pharmacoepidemiol. Drug Safety*, vol. 20, no. 3, pp. 292–299, 2011.
[19] I. Zorych, D. Madigan, P. Ryan, and A. Bate, "Disproportionality methods for pharmacovigilance in longitudinal observational databases," *Statist. Methods Med. Res.*, vol. 22, no. 1, pp. 39–56, 2013.
[20] 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 *Advances in Knowledge Discovery and Data Mining*. New York, NY, USA: Springer, 2006, pp. 867–876.
[21] 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 Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 839–853, Jun. 2010.
[22] J. M. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. E. Gibson, and R. B. Hubbard. (2013). Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Comput.* [Online]. Available: http://link.springer.com/content/pdf/10.1007%2Fs00500-013-1097-4.pdf
[23] 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," *Pharmacoepidemiol. Drug Safety*, vol. 14, no. 7, pp. 443–451, 2005.
[24] NetDoctor. (2012). *NetDoctor.co.uk—the UK's leading independent health website.* [Online]. Available: http://www.netdoctor.co.uk
[25] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 1–26, 2012.
[26] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, vol. 2, pp. 27–34.
[27] D. J. Graham, D. Campen, R. Hui, M. Spence, C. Cheetham, G. Levy, S. Shoor, and W. A. Ray, "Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: Nested case-control study," *Lancet*, vol. 365, no. 9458, pp. 475–481, 2005.
[28] J.-P. Després, A. Golay, and L. Sjöström, "Effects of rimonabant on metabolic risk factors in overweight patients with dyslipidemia," *New England J. Med.*, vol. 353, no. 20, pp. 2121–2134, 2005.
[29] R. Christensen, P. K. Kristensen, E. M. Bartels, H. Bliddal, and A. Astrup, "Efficacy and safety of the weight-loss drug rimonabant: A meta-analysis of randomised trials," *Lancet*, vol. 370, no. 9600, pp. 1706–1713, 2007.
[30] A. Carnovale, P. Esposito, P. Bassano, L. Russo, and G. Uomo, "Enalapril-induced acute recurrent pancreatitis," *Digest. Liver Disease*, vol. 35, no. 1, pp. 55–57, 2003.
[31] A. R. Balani and J. H. Grendell, "Drug-induced pancreatitis," *Drug Safety*, vol. 31, no. 10, pp. 823–837, 2008.
[32] S. Harirforoosh and F. Jamali, "Renal adverse effects of nonsteroidal anti-inflammatory drugs," *Expert Opin. Drug Safety*, vol. 8, no. 6, pp. 669–681, 2009.
[33] S. Ali, J. D. Pimentel, and C. Ma, "Naproxen-induced liver injury," *Hepatobiliary Pancreat. Diseases Int.*, vol. 10, no. 5, pp. 552–556, 2011.
[34] G. Onder, F. Pellicciotti, G. Gambassi, and R. Bernabei, "NSAID-related psychiatric adverse events," *Drugs*, vol. 64, no. 23, pp. 2619–2627, 2004.
[35] Drugs.com. (2012). *Nifedipine side effects from Drugs.com.* [Online]. Available: http://www.drugs.com/sfx/nifedipine-side-effects.html
[36] P. D. Van der Linden, M. C. J. M. Sturkenboom, R. M. C. Herings, H. G. M. Leufkens, and B. H. Stricker, "Fluoroquinolones and risk of Achilles tendon disorders: Case-control study," *Brit. Med. J.*, vol. 324, no. 7349, pp. 1306–1307, 2002.
[37] C. H. Lee and C. J. Lee, "Thrombocytopenia—A rare but potentially serious side effect of initial daily and interrupted use of rifampicin," *Chest J.*, vol. 96, no. 1, pp. 202–203, 1989.
[38] N. N. Chan, H. P. S. Brain, and M. D. Feher, "Metformin-associated lactic acidosis: A rare or very rare clinical entity?" *Diabet. Med.*, vol. 16, no. 4, pp. 273–281, 1999.
[39] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, "Web-scale pharmacovigilance: Listening to signals from the crowd," *J. Amer. Med. Inf. Assoc.*, vol. 20, no. 3, pp. 404–408, 2013.
[40] 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," *Statist. Med.*, vol. 31, no. 30, pp. 4401–4415, 2012.

Authors' photographs and biographies not available at the time of publication.