Temporal Weighting of Clinical Events in Electronic Health Records for Pharmacovigilance

Jing Zhao
Department of Computer and Systems Sciences (DSV)
Stockholm University
Stockholm, Sweden
Email: jingzhao@dsv.su.se

***Abstract*—Electronic health records (EHRs) have recently been identified as a potentially valuable source for monitoring adverse drug events (ADEs). However, ADEs are heavily under-reported in EHRs. Using machine learning algorithms to automatically detect patients that should have had ADEs reported in their health records is an efficient and effective solution. One of the challenges to that end is how to take into account temporality when using clinical events, which are time stamped in EHRs, as features for machine learning algorithms to exploit. Previous research on this topic suggests that representing EHR data as a bag of temporally weighted clinical events is promising; however, how to assign weights in an optimal manner remains unexplored. In this study, nine different temporal weighting strategies are proposed and evaluated using data extracted from a Swedish EHR database, where the predictive performance of models constructed with the random forest learning algorithm is compared. Moreover, variable importance is analyzed to obtain a deeper understanding as to why a certain weighting strategy is favored over another, as well as which clinical events undergo the biggest changes in importance with the various weighting strategies. The results show that the choice of weighting strategy has a significant impact on the predictive performance for ADE detection, and that the best choice of weighting strategy depends on the target ADE and, specifically, on its dose-dependency.**

I. INTRODUCTION

Adverse drug events (ADEs), or drug side effects, are often defined as undesired harms resulting from the use of a drug and cause approximately 2.4% to 12.0% of hospital admissions worldwide [1]–[3]. As a result, they are considered to constitute a major public health problem. Every year, many drugs are withdrawn from the market due to their severe, hitherto unknown ADEs. Examples include Vioxx for its doubled risk of causing myocardial infarction [4] and Cerivastatin for causing fatal rhabdomyolysis [5]. Moreover, most ADEs, in particular those that are not dose-dependent, result from inappropriate prescriptions of drugs and are therefore preventable [6]. Pharmacovigilance is a research area that aims to improve drug safety, pre- and post-marketing, primarily using resources such as clinical trials, spontaneous reports and longitudinal healthcare databases [7]. However, due to the limitations of clinical trials, in terms of number of participants and follow-up time, and of spontaneous reports, in terms of reporting rate and reliability [8], electronic health records (EHRs) have recently emerged as a potentially valuable source for pharmacovigilance [9]–[11].

Electronic health records have several advantages over traditional sources of information for drug safety surveillance: (1) they contain longitudinal healthcare data over a long time period across a large population; (2) EHR data provides a holistic perspective of patient health history, including diagnoses, drug admissions, laboratory tests, etc.; and (3) this data is more reliable since it is reported by clinical professionals in the real clinical setting. Nevertheless, ADEs are still under-reported in EHRs [12]. Manually screening millions of health records to identify ADEs is practically impossible for the massive amounts of data archived in an EHR database. To mitigate this problem, supervised machine learning can be adopted to automatically detect the presence of an ADE in health records in which it was not but should have been reported [13]–[19]. To that end, predictive models are trained to detect health records that contain ADEs with clinical events – i.e., diagnoses, drugs, clinical measurements, etc. – as features. These clinical events are reported in a chronological order in EHRs and the same event often appears in the same health record several times at different time points. How to handle the temporality of clinical events in the context of using supervised machine learning for ADE detection remains a challenge.

Supervised machine learning algorithms learn from features or predictors that describe training examples in order to find patterns that can distinguish examples that belong to different classes. If clinical events in EHRs are used as distinct features, an important task is then to represent these chronological events in a way that temporality is taken into account in a manner that leads to the best possible predictive performance. In a previous study, which focuses on representations of clinical events, two methods that handle temporality of clinical events in EHRs were proposed and evaluated [20], where the first one treats the same event that occurred at different time points as different features and the second one assigns different weights to the same event that occurred at different time points and then aggregates them. The former creates additional features according to temporality, i.e., each unique event is transformed into multiple bins that cover a certain time period relative to time distance from the target ADE; the latter, instead, injects temporality into the calculation of feature values. It was shown that both methods lead to better predictive performance than simply ignoring the temporality – by modeling the data as a bag of (unweighted) clinical events – and that the second method yields the best results in general. However, only a single weighting strategy was considered in that study and no alternatives were evaluated that would suggest that the chosen one is the most optimal.

This study aims to explore various temporal weighting

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

strategies and their impact on the performance of predictive models for ADE detection in EHRs. Figure 1 illustrates how weighted clinical events can be extracted from EHRs, which will subsequently be used as features by supervised machine learning algorithms; an example of a health record spanning three days and comprising three different types of clinical events (drug, diagnosis and measurement) is shown. To study this problem thoroughly, nine different weighting strategies are proposed in this study, and they are evaluated by experimenting on several ADEs with corresponding health records that are extracted from a real Swedish EHR database. The use case is to distinguish health records in which an ADE has been reported from health records in which a similar diagnosis code – but one that does not indicate that it has been caused by drugs – has been assigned.

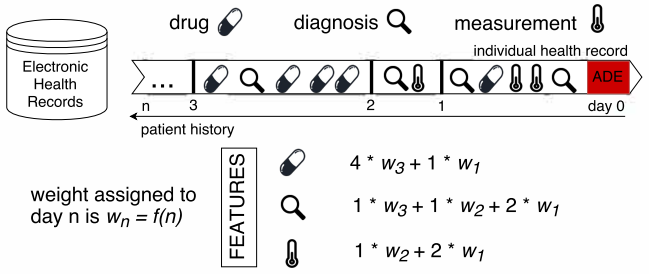

Fig. 1. Extract weighted clinical events from electronic health records

II. METHODS AND MATERIALS

Here, the nine proposed temporal weighting strategies, along with their underlying assumption, are first introduced. To evaluate their impact on predictive performance, a series of experiments are designed using random forest as the supervised machine learning algorithm and conducted on fourteen datasets, corresponding to fourteen ADEs, that are extracted from Stockholm EPR Corpus, a Swedish EHR database. For each patient, a health record of 90 days before the target ADE is analyzed. To assess the predictive performance of random forest models using different temporal weighting strategies, area under ROC curve is used as the main performance evaluation metric. Finally, variable importance is analyzed in order to gain more evidence on which clinical events are influenced most by the choice of weighting strategy.

A. Temporal weighting Strategies

In this study, a temporal weighting strategy follows the following form: for a clinical event that occurred n days prior to the occurrence of an ADE, weight w is assigned according to a curve function $f(n)$. The common underlying assumption for these nine strategies is that events that occurred closer to the target ADE are more important, in terms of their informativeness in the predictive models, and should therefore receive more weight than those that occurred a longer time before the target ADE. Therefore, events that occurred in the same day as the target ADE receive a weight of 1, the highest weight, and then the weight decreases monotonically with increasing number of days between the corresponding event and the target ADE; since the patient history is limited to 90 days[1] in this study, the weight for events that occurred 91 days, and more, before the target ADE is 0 (note that one day is added here to make sure that events on the 90th day receive low but non-zero weights); all assigned weights are between 0 and 1. In such a situation, the nine temporal weighting strategies differ in the speed in which weights decrease along the patient history, as illustrated in Figure 2. In this figure, $w1$ to $w9$ are shown from left to right, ordered by the strength of each strategy, where the impact of strategies to the left is stronger, i.e., the weights drop quicker, than the ones to the right.

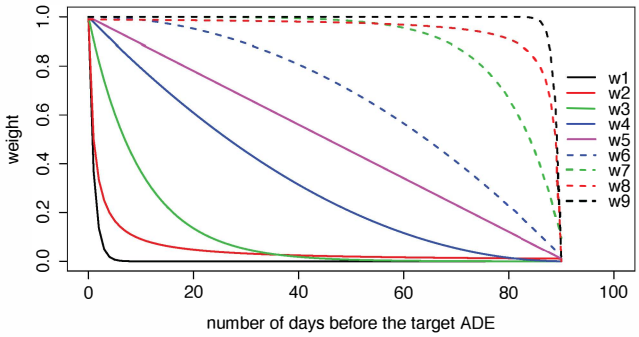

Fig. 2. Nine weighting strategies following nine curve functions

In $w1 - w4$, each weighting strategy is relatively harsh in a way that the weight starts from 1 on day 0 and then immediately decreases sharply within a number of days (varies across different strategies), which indicates that clinical events that occurred a few days before the target ADE are assumed to be much more relevant than earlier events.

- $w1$ follows an exponential function of n:

$$w1 = exp(-n)$$

- $w2$ follows a reciprocal function of n:

$$w2 = \frac{1}{n+1}$$

- $w3$ also follows an exponential function of n, but much softer than $w1$:

$$w3 = exp(-0.1 \times n)$$

- $w4$ follows a second degree polynomial function of n:

$$w4 = \frac{(n-91)^2}{91^2}$$

In $w5$, the weighting strategy assumes that the importance of a clinical event is directly proportional to the number of days between it and the target ADE.

- $w5$ follows a linear function of n:

$$w5 = 1 - \frac{n}{91}$$

In $w6 - w9$, each weighting strategy is relatively soft compared to $w1 - w5$, where the weight starts from 1 on

[1]"90 days" was chosen arbitrarily with common sense, i.e., the drugs or other clinical events that occurred more than 3 months before the occurrence of an ADE were considered with no significant contribution.

day 0 but it only decreases softly until a number of days (varies across strategies) before the beginning of patient history – 90 days in this case – and then decreases towards 0 quickly. Such strategies indicate that clinical events that occurred much earlier than the target ADE are also relevant indicators, though not as much as the immediate events.

- $w6$ follows a second degree polynomial function of n, asymmetrical to $w4$:

$$w6 = 1 - \frac{n^2}{91^2}$$

- $w7$ follows a transformed exponential function of n, asymmetrical to $w3$:

$$w7 = 1 - exp(0.1 \times (n - 91))$$

- $w8$ follows a transformed reciprocal function of n, asymmetrical to $w2$:

$$w8 = 1 - \frac{1}{91 - n}$$

- $w9$ follows a transformed exponential function of n, asymmetrical to $w1$:

$$w9 = 1 - exp(n - 91)$$

B. Data Source

In this study, 14 datasets were extracted from a real EHR system – the Stockholm EPR Corpus[2] [21]. This database contains health records from around 700,000 patients over two years (2009-2010), which were collected from Karolinska University Hospital in Stockholm, Sweden. Various types of information are available in this database for each patient, including 10,000 unique diagnoses (encoded by ICD-10[3]), 1,500 unique drugs (encoded by ATC[4]), 730 unique clinical measurements and millions of clinical notes in free-text. Here, we only used the structured information, i.e., diagnoses, drugs and clinical measurements.

An earlier study [22] has categorized ICD-10 diagnosis codes in terms of how they are used for indicating ADEs during hospital admissions, among which category A.1 (a drug-related causation was noted in the diagnosis code) and category A.2 (a drug- or other substance-related causation was noted in the diagnosis code) indicate a clear sign of ADE occurrence; hence the most frequent A.1 and A.2 ADE-related diagnosis codes in the Stockholm EPR Corpus were selected. In total, 14 datasets were created with the existence of an ADE-related diagnosis code as the class label in each dataset. The task here is to detect patients who should, but do not, have a specific ADE reported in their health records, which results in a binary classification task.

Examples – positive examples were patients whom have been assigned an ADE-specific diagnosis code and negative examples were patients whom have been assigned a similar code (defined as two codes sharing the same first three levels of the ICD-10 concept hierarchy) to the corresponding ADE-related one. For instance, if the positive examples were patients diagnosed with *G44.4* (drug-induced headache), the negative examples were patients diagnosed with any code starting with *G44* (other headache syndromes), but not *G44.4*.

Features – unique clinical events including diagnoses, drugs and clinical measurements, that occurred 90 days before the occurrence of the target ADE were used as features. According to the findings from an earlier study on representing clinical events in EHRs [13], for each example, the value for each clinical event (feature) was the *total number of times* that it occurred in the patient history up to 90 days. The occurrence of each unique event at different time point was weighted according to one of the weighting strategies and then these weighted occurrences were summed up to obtain the "weighted number of times", as illustrated in Figure 1.

All datasets in this study are of high dimensionality and sparsity due to the fact that most clinical events only occurred to a small group of patients, i.e., the vast majority of the examples for a given feature have a value of zero. Therefore, features that are more sparse than 99%, i.e., the ones for which non-zero values were observed in fewer than 1% of the examples, were removed; for those datasets with fewer than one hundred observations, features with only one non-zero value were also removed. The motivation for this is two-fold: (1) to reduce the dimensionality and sparsity; (2) to highlight the impact of applying different weighting strategies. The former is intuitive; the latter is motivated by the fact that for features with only one or very few non-zero values, the impact of applying different weighting strategies is almost negligible, even though some of these features might be valuable indicators. To be more specific, if a feature has only one or a few non-zero values, when using it as an indicator to classify the examples, it does not matter whether these small numbers of non-zero values are weighted or not since they will most likely be distinguished against all the zero values; as a result, a weighting strategy will almost have no impact on the predictive performance. Table I lists basic descriptions of each dataset, including the diagnosis code that indicates the corresponding ADE (dataset name), the description of this code, the number of positive and negative examples, the number of features and amount of sparsity (from both the original datasets and the reduced ones).

C. Experimental Setup

In this study, two consecutive experiments were conducted to evaluate the proposed temporal weighting strategies in terms of their impact on predictive performance when detecting ADEs. In the first experiment, nine weighting strategies – $w1$ to $w9$ – were applied to the clinical events that were sorted chronologically in each patient's health record to generate the corresponding features in each dataset, which were then fitted by the random forest algorithm [23] to generate predictive models. Random forest was chosen mainly for its reputation of being robust in terms of achieving high accuracy, its ability to handle high-dimensional data efficiently, as well as the possibility of obtaining estimates of variable importance. This algorithm is an ensemble classifier, which constructs a set of decision trees together voting for what class label to assign to

[2]This research has been approved by the Regional Ethical Review Board in Stockholm (permission number 2012/834-31/5).

[3]The 10th revision of the International Statistical Classification of Diseases and Related Health Problems

[4]Anatomical Therapeutic Chemical Classification System

TABLE I. DATASETS DESCRIPTION

Dataset	Corresponding diagnosis code description	No. of Examples		No. of Features		Sparsity (%)	
		Positive	Negative	Original	Reduced	Original	Reduced
D642	Secondary sideroblastic anemia due to drugs and toxins	113	4234	3970	381	99.33	94.33
G240	Drug-induced dystonia	16	44	408	408	97.91	97.91
G444	Drug-induced headache, not elsewhere classified	31	1102	1370	103	99.42	94.48
G620	Drug-induced polyneuropathy	19	367	1444	223	99.05	95.62
I952	Hypotension due to drugs	38	480	1538	366	98.53	94.78
L270	Generalized skin eruption due to drugs and medicaments	174	291	1305	268	98.77	95.34
L271	Localized skin eruption due to drugs and medicaments	58	407	1311	266	98.78	95.33
O355	Maternal care for (suspected) damage to fetus by drugs	334	373	628	130	98.90	95.63
T782	Adverse effects: anaphylactic shock, unspecified	136	1467	1383	107	99.47	94.85
T783	Adverse effects: angioneurotic oedema	147	1448	1383	110	99.47	94.95
T784	Adverse effects: allergy, unspecified	984	612	1379	110	99.46	94.95
T808	Other complications following infusion, transfusion and therapeutic injection	353	59	1268	325	97.86	92.79
T886	Anaphylactic shock due to correct drug or medicament properly administered	53	607	2123	322	98.98	94.66
T887	Unspecified adverse effect of drug or medicament	472	277	2129	324	98.97	94.66

an example to be classified. Each tree in the forest is built from a bootstrap replicate of the original instances, and a subset of all features is randomly sampled at each node when building the tree, in both cases to increase diversity among the trees. With increasing number of trees in the forest, the probability that a majority of trees makes an error decreases, given that the trees perform better than random and that the errors are made independently. The algorithm has often been shown in practice to result in state-of-the-art predictive performance, though this condition can only be guaranteed in theory. In this study, random forest was implemented with 500 trees.

The generated predictive models were evaluated via stratified 5-fold cross validation with 10 iterations. The primary performance evaluation metric was area under the ROC curve (AUC), which depicts the performance of a model without regard to class distribution or error costs by estimating the probability that a model ranks a randomly chosen positive example ahead of a negative one. AUC is preferred here due to the unbalanced class distribution in each dataset, to which AUC is not biased towards. Other commonly used performance metrics – accuracy, precision, recall, F_1-score and area under the precision-recall curve (AUPRC) – were also reported in this study to evaluate the results from various perspectives. Accuracy calculates the percentage of examples that are correctly classified. Precision measures the fraction of true positives among all the predicted positives, while recall, also known as sensitivity, measures the fraction of true positives among all the positives in the reference standard. In the case of detecting ADEs, high precision means that the algorithm is able to detect more true ADEs than false ones, while high recall indicates the capacity of detecting most true ADEs. F_1-score is the harmonic mean of precision and recall by calculating $2 \times (precision \times recall)/(precision + recall)$. Only both a high precision and a high recall can yield a high F_1-score. At last, AUPRC depicts the probability that precision is higher than recall for each recall threshold.

The Friedman test [24] was employed for statistical testing of the null hypothesis that all models perform equally, i.e., that the choice of weighting strategy has no impact on the predictive performance. The nine proposed weighting strategies were not only compared to each other, but also to a baseline strategy, $w0$, where no weighting is involved.

- $w0$ assigns every clinical event a weight of 1:

$$w0 = 1, \forall n$$

In a follow-up experiment, variable importance generated from the random forest models using the best and worst weighting strategy, as observed in the previous experiment, was analyzed to obtain a deeper understanding of the differences between them. Variable importance can be estimated in different ways, see, e.g., [23]. In this study, Gini importance [25] was used as the variable importance metric, where a high Gini importance indicates that a variable plays a greater role in splitting the data into the defined classes. A Gini importance of zero means that a variable is considered useless or is never selected to build any tree in the forest. Here, variable importance was analyzed on two levels: (1) global level – ranking the difference in features' Gini importance in general; (2) local level – specific features whose Gini importance rank changes most dramatically between the two chosen strategies.

III. RESULTS

To compare the predictive performance of random forest models using different weighting strategies, they were ranked based on the chosen performance evaluation metrics, respectively, for each dataset. The averaged ranks of each weighting strategy over 14 datasets are presented in Table II (note that the ranks here range from 1 to 10 given there are 10 weighting strategies to compare and higher ranks indicate worse performance), from which we can see that $w2$ yields the best results with all metrics, while the baseline, $w0$, is the worst with five out of six metrics. The impact of different weighting strategies on the predictive performance is significant with all metrics but recall.

TABLE II. AVERAGED RANKS OF PREDICTIVE PERFORMANCE FROM RANDOM FOREST MODELS USING DIFFERENT WEIGHTING STRATEGIES. (NUMBER IN BOLD INDICATES THE BEST AND ASTERISK THE WORST)

Strategy	Accuracy	AUC	AUPRC	Precision	Recall	F_1-score
$w1$	6.96	7.14	4.57	5.28	5.53	5.93
$w2$	**3.14**	**3.86**	**3.29**	**3.96**	**4.11**	**3.43**
$w3$	3.82	4.64	4.00	4.25	4.57	4.00
$w4$	5.64	4.92	6.42	5.68	5.61	5.82
$w5$	5.46	5.71	5.71	5.17	5.64	5.68
$w6$	5.29	4.14	5.21	5.14	5.17	4.79
$w7$	4.82	4.93	5.64	5.18	5.64	5.54
$w8$	5.46	4.57	5.57	5.68	5.75	5.96
$w9$	7.25*	7.14	7.00	7.21	6.11	6.36
$w0$	7.14	7.93*	7.57*	7.43*	6.86*	7.50*
P-value	0.002	0.001	0.004	0.020	0.395	0.009

Adverse drug events differ from each other in terms of dose-dependency, i.e., the negative effect of a drug depends on levels of exposure after a certain exposure time, which, in fact,

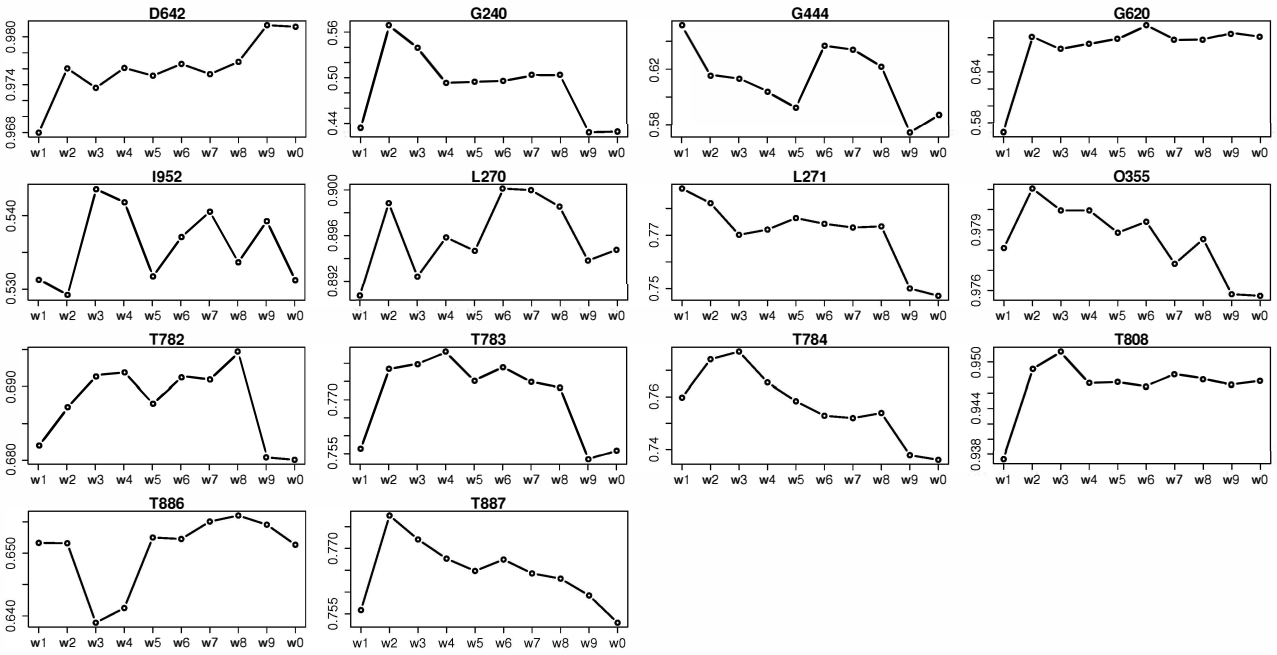

Fig. 3. AUC scores for each dataset obtained from random forest models using the nine weighting strategies ($w1$ – $w9$) and the baseline (no weighting, $w0$)

has a direct connection with the choice of weighting strategy. Intuitively, for ADEs that are highly dose-dependent, clinical events that occurred a while ago should not be much less important than the more immediate events; for the ones that are not dose-dependent, however, only the immediate events are important contributors. Figure 3 shows the AUC scores that are obtained from random forest models using all weighting strategies ($w1 - w0$) for each dataset (ADE), respectively. For some ADEs, such as *L271*, it is obvious that a harsher weighting strategy yields better AUC, while for *D642* the result is completely the opposite, with almost no weighting yielding the best result.

In the second experiment – variable importance analysis – the observed best strategy from the previous experiment, $w2$, was compared to the worst, $w0$ (no weights assigned). Features were ranked according to their Gini importance and the rank difference of each feature between $w2$ and $w0$ was calculated. In Table III, the results of the global level analysis is shown, which include, for each dataset, the number of features that are higher ranked in models using $w2$ and $w0$, respectively, and the average (absolute) rank difference of all features. On the one hand, some datasets have more features ranked higher in $w2$, while some in $w0$, and the number of features on each side is fairly close to each other; compared to total number of features in each dataset, almost all features are ranked differently between these two weighting strategies[5]. On the other hand, the average rank difference, to some extent, reflects the predictive performance difference between models using $w2$ and $w0$ (see Figure 3), i.e., higher average rank difference corresponds to bigger difference in predictive performance.

[5]Note that the total number of features that are ranked differently in the two weighting strategies is sometimes larger than the total number of (reduced) features shown in Table I. This is because that the total number of features is averaged over 5-fold cross validation, while in the variable importance analysis, all features that are ranked differently from each fold are counted.

TABLE III. GLOBAL LEVEL VARIABLE IMPORTANCE ANALYSIS

Dataset	Higher ranked in $w2$	Higher ranked in $w0$	Average rank difference
D642	214	209	19.66
G240	48	44	6.85
G444	59	68	9.17
G620	134	123	15.40
I952	206	192	21.37
L270	139	162	11.77
L271	158	146	17.90
O355	73	68	7.22
T782	66	62	8.01
T783	59	70	5.97
T784	65	66	6.51
T808	184	194	16.93
T886	203	193	24.96
T887	193	200	18.82

In the local level variable importance analysis, features were ranked according to their rank differences between models using $w2$ and $w0$, and the top 5 highest ranked features for each dataset are listed in Table IV. Note that these features are not necessarily the most informative / important, but the ones whose importance changes the most dramatically between $w2$ and $w0$. Among the listed clinical events, some are relevant indicators of the corresponding ADE, while others seem to constitute random noise. Here, a few examples (made bold in Table IV) are picked to illustrate how such results can be understood. Symptoms like dizziness and abdominal pain are often immediate effects that indicate the presence of an ADE; therefore they are more informative when only the ones that occurred close to the target ADE are considered important in the model. Clinical events such as type 1 and type 2 diabetes are typically ranked higher in models using $w0$, since they are chronic diseases and thus their impact over time should be more or less constant. All allergy-related features are ranked higher in models using $w0$, which indicates that such features are more informative when the model does not assign lower weights to them even if they occurred a while ago. This makes

sense as we know that, in fact, a patient's allergy information should be an important indicator, even though the reporting time was a while ago.

TABLE IV. LOCAL LEVEL VARIABLE IMPORTANCE ANALYSIS. FEATURE NAME (RANK DIFFERENCE). POSITIVE RANK DIFFERENCE INDICATES FEATURE RANKED HIGHER IN W2, NEGATIVE W0

Dataset	Top 5 features ranked most differently between w0 and w2
D642	Ostomy, liquid amount (136)
	Dizziness and vertigo (134)
	Atrial fibrillation and atrial flutter, unspecified (-101)
	Unspecified malignant tumor of the pancreas (-97)
	Leg ulcers, not elsewhere classified (89)
G240	Tramadol (-25)
	Docusate sodium and e.g. sorbitol or glycerol (24)
	Rivaroxaban (22)
	Swallowing difficulties (-21)
	P-Glucose (-21)
G444	Asthma, unspecified (-41)
	Cyanocobalamin (-39)
	Counseling, unspecified (34)
	Sodium picosulfate (34)
	Observation for suspected diseases and conditions (-29)
G620	Felodipine (95)
	Atrial fibrillation and atrial flutter, unspecified (79)
	Counseling, unspecified (60)
	Unspecified malignant tumor in the mammary gland (58)
	Presence of electronic cardiac device (58)
I952	Adjustment and management of cardiac device (-169)
	U volume in 24h (-133)
	midazolam (117)
	Pain or aches, unspecified (-111)
	Ordinary salt combinations (99)
L270	Thrombocytopenia, unspecified (-68)
	Head circumference (54)
	Granisetron (-49)
	Walking ability, according to Downton (48)
	Allergic urticaria (-47)
L271	Dermatitis, unspecified (-118)
	Targeted health check regarding atopic disease (-97)
	Counseling, unspecified (90)
	Allergy, unspecified (-81)
	Observation for suspected diseases and conditions (-73)
O355	**Type 1 diabetes mellitus before pregnancy (-41)**
	Person with feared disease in which no diagnosis (-40)
	nitrofurantoin (-34)
	Supervision of other normal pregnancy (-29)
	Body temperature (-27)
T782	Urticaria, unspecified (-44)
	FEV1 (-37)
	Patch test Birch (26)
	Cow's milk allergy (-23)
	Prick test: timothy (23)
T783	Constipation (37)
	Ketobemidon (-20)
	Macrogol, combinations (-20)
	B-CRP (19)
	Prick test: horse (19)
T784	Dermatitis caused by ingested food (-33)
	Karbamid (-27)
	Prick test: timothy (-24)
	Renewal of recipes (23)
	FEV1 (-22)
T808	Magnesium oxide (96)
	Amoxicillin and enzyme inhibitor (92)
	Iopromide (85)
	Dextropropoxyphene (79)
	Fentanyl, combinations (78)
T886	Pneumonia, unspecified (159)
	Abdominal pain (134)
	Local anesthetics, combinations (120)
	Secondary malignant tumor of the lymph nodes (-113)
	Residual (-106)
T887	**Type 2 diabetes mellitus (-93)**
	Other specified counseling (-91)
	Carbamide (-90)
	Flucloxacillin (78)
	Urination frequency (77)

IV. DISCUSSION

In longitudinal healthcare databases, such as EHRs, clinical events are often time stamped. Temporal information is valuable when making use of such sources for pharmacovigilance. On the other hand, it is also challenging to take into account the temporality if we would like to fit such data into supervised machine learning models for ADE detection. This study investigated this problem from the angle of how to embed temporality into feature representations that can be used by predictive models and, particularly, how to assign weights to clinical events that occurred at different time points.

Here, nine temporal weighting strategies were proposed and evaluated, where the weights are assigned according to the temporal relationship between clinical events and the target ADE following different curve functions. These weighted clinical events are then summed for each unique event as features to be used by the random forest learning algorithm to detect patient health records that have recorded ADEs. The predictive performance – AUC scores of 0.7 to 0.9 for most datasets – demonstrates the effectiveness of the applied method. By comparing the weighting strategies to the baseline, it is clear that assigning temporal weights to clinical events leads to better predictive performance than no weighting at all; among the nine strategies, the one that follows a reciprocal function, $w2$, yields the best result. This entails that events that occurred a long time before the target ADE receive very low weights, i.e., their existence should only be taken into account to a limited extent, compared to events that occurred in close proximity to the ADE. Furthermore, given the poor performance of models using $w1$, which is a very harsh weighting strategy, in the way that weights reduce sharply immediately after one or two days before the target ADE, we can conclude that events in close temporal proximity to the target ADE should not take such a strong precedence over earlier events (see Figure 2 for relationships between different weighting strategies).

For each ADE that was investigated in this study, the best weighting strategy is not always the same. For instance, *D642 (drug induced anemia)* favors the very mild strategy, with almost no weighting, which indicates that to detect such an ADE, historical clinical events of a patient are very important indicators; for ADEs like *L271 (drug induced skin eruptions)* or *O355 (drug induced damage to fetus)*, on the other hand, the harsher strategies, which only consider events close to the target ADE to be important, result in better predictive performance. This is not particularly surprising, as when patient history is used for detecting ADEs, their dose-dependency is a very important component in the decision-making. Similarly, in a real-life setting, physicians will ask about their patients' recent clinical activities if they suffer from a dose-independent ADE, but they will want to know what happened for, e.g., the last three months if they suspect dose-dependent ones. However, the dose-dependency of an ADE is not always obvious; therefore it is still interesting to find out which weighting strategy should be adopted in such a situation.

To gain a deeper understanding on the impact of different weighting strategies, variable importance, obtained from the random forest learning algorithm using the best strategy ($w2$) and the baseline ($w0$), also the worst, was analyzed. Among events whose importance changes dramatically between these

two weighting strategies, there are also some seemingly irrelevant ones; therefore, domain experts are perhaps still needed to filter out some of the irrelevant events if we ought to use such methods for prediction tasks in a real clinical setting.

One limitation of this study is that weights are pre-assigned to all clinical events and not obtained through learning from the data, which consequently limits the precision of the weighting strategy and also disallows a tailor-made weighting strategy for each ADE. Moreover, in this study, weights are assumed to decrease monotonically, albeit at different speeds, along the patient history, which seems rather reasonable; however, alternatives have not been explored. For instance, a third degree polynomial function could be an alternative here. This is also related to the previous point that such pre-assigned weights are limited in capturing specific characteristics of each feature. For future work, it would be interesting to explore how to learn the weights from data, instead of pre-assigning weights. Another limitation of this study is that everything that occurred within 90 days of patient history is included in the predictive model, which at the same time introduces a risk of noise; the results, especially from the variable importance analysis, would be more precise and relevant if a clinical expert filtered out some obviously irrelevant events prior to the learning process.

V. CONCLUSION

This study tackles the problem of making the best use of time stamped clinical events in electronic health records by means of supervised machine learning for detecting the presence of a particular adverse drug event in patient health records. Particularly, the focus here is on how to assign weights to these events in accordance with their temporal relationships to the target adverse drug event. It is concluded that the choice of weighting strategy has a significant impact on the predictive performance, and that the dose-dependency of an adverse drug event should be taken into account in the decision-making.

ACKNOWLEDGMENT

This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection, ref. no. IIS11-0053.

REFERENCES

[1] S. Schneeweiss, J. Hasford, M. Göttler, A. Hoffmann, A.-K. Riethling, and J. Avorn, "Admissions caused by adverse drug events to internal medicine and emergency departments in hospitals: a longitudinal population-based study," *European Journal of Clinical Pharmacology*, vol. 58, no. 4, pp. 285–291, 2002.

[2] 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," *BMJ*, vol. 329, no. 7456, pp. 15–19, 2004.

[3] T. Mjörndal, M. D. Boman, S. Hägg, M. Bäckström, B.-E. Wiholm, A. Wahlin, and R. Dahlqvist, "Adverse drug reactions as a cause for admissions to a department of internal medicine," *Pharmacoepidemiology and Drug Safety*, vol. 11, no. 1, pp. 65–72, 2002.

[4] B. Sibbald, "Rofecoxib (vioxx) voluntarily withdrawn from market," *Canadian Medical Association Journal*, vol. 171, no. 9, pp. 1027–1028, 2004.

[5] C. D. Furberg and B. Pitt, "Withdrawal of cerivastatin from the world market," *Curr Control Trials Cardiovasc Med*, vol. 2, no. 5, pp. 205–207, 2001.

[6] H. Lövborg, L. R. Eriksson, A. K. Jönsson, T. Bradley, and S. Hägg, "A prospective analysis of the preventability of adverse drug reactions reported in sweden," *European Journal of Clinical Pharmacology*, vol. 68, no. 8, pp. 1183–1189, 2012.

[7] L. Härmark and A. Van Grootheest, "Pharmacovigilance: methods, recent developments and future perspectives," *European Journal of Clinical Pharmacology*, vol. 64, no. 8, pp. 743–752, 2008.

[8] S. A. Goldman, "Limitations and strengths of spontaneous reports data," *Clinical Therapeutics*, vol. 20, pp. C40–C44, 1998.

[9] M. J. Schuemie, P. M. Coloma, H. Straatman, R. M. Herings, G. Trifirò, J. N. Matthews, D. Prieto-Merino, M. Molokhia, L. Pedersen, R. Gini *et al.*, "Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods," *Medical Care*, vol. 50, no. 10, pp. 890–897, 2012.

[10] I. Karlsson, J. Zhao, L. Asker, and H. Boström, "Predicting adverse drug events by analyzing electronic patient records," in *Artificial Intelligence in Medicine Lecture Notes in Computer Science*. Springer, 2013, pp. 125–129.

[11] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman, "Mining electronic health records for adverse drug effects using regression based methods," in *1st ACM International Health Informatics Symposium*. ACM, 2010, pp. 100–107.

[12] L. Hazell and S. A. Shakir, "Under-reporting of adverse drug reactions," *Drug Safety*, vol. 29, no. 5, pp. 385–396, 2006.

[13] J. Zhao, A. Henriksson, L. Asker, and H. Boström, "Detecting adverse drug events with multiple representations of clinical measurements," in *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2014, pp. 536–543.

[14] J. Zhao, A. Henriksson, and H. Boström, "Detecting adverse drug events using concept hierarchies of clinical codes," in *IEEE International Conference on Healthcare Informatics*. IEEE, 2014, pp. 285–293.

[15] A. Henriksson, J. Zhao, H. Boström, and H. Dalianis, "Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection," in *IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2015.

[16] J. Zhao, A. Henriksson, and H. Boström, "Cascading adverse drug event detection in electronic health records," in *IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2015.

[17] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *Journal of Biomedical Informatics*, vol. 57, pp. 333–349, 2015.

[18] A. Henriksson, "Representing clinical notes for adverse drug event detection," in *6th International Workshop on Health Text Mining and Information Analysis (LOUHI)*. ACL, 2015, pp. 152–158.

[19] A. Henriksson, J. Zhao, H. Boström, and H. Dalianis, "Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection," in *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2015.

[20] J. Zhao, A. Henriksson, M. Kvist, L. Asker, and H. Boström, "Handling temporality of clinical events for drug safety surveillance," in *American Medical Informatics Association (AMIA) Annual Symposium*, 2015.

[21] H. Dalianis, M. Hassel, A. Henriksson, and M. Skeppstedt, "Stockholm epr corpus: a clinical database used to improve health care," in *Swedish Language Technology Conference*, 2012.

[22] J. Stausberg and J. Hasford, "Drug-related admissions and hospital-acquired adverse drug events in germany: a longitudinal analysis from 2003 to 2007 of icd-10-coded routine data," *BMC Health Services Research*, vol. 11, no. 1, p. 134, 2011.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[25] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.