Handling Temporality of Clinical Events for Drug Safety Surveillance

Jing Zhao¹, MS, Aron Henriksson¹, MS, Maria Kvist¹,², PhD, MD, Lars Asker¹, PhD, Henrik Boström¹, PhD
¹ Dept. of Computer and Systems Sciences, Stockholm University, Sweden
² Dept. of Learning, Informatics, Management and Ethics, Karolinska Institute, Sweden

Abstract

Using longitudinal data in electronic health records (EHRs) for post-marketing adverse drug event (ADE) detection allows for monitoring patients throughout their medical history. Machine learning methods have been shown to be efficient and effective in screening health records and detecting ADEs. How best to exploit historical data, as encoded by clinical events in EHRs is, however, not very well understood. In this study, three strategies for handling temporality of clinical events are proposed and evaluated using an EHR database from Stockholm, Sweden. The random forest learning algorithm is applied to predict fourteen ADEs using clinical events collected from different lengths of patient history. The results show that, in general, including longer patient history leads to improved predictive performance, and that assigning weights to events according to time distance from the ADE yields the biggest improvement.

Introduction and Motivation

Drug safety is a critical public health issue, with the prevalence of hospital admissions caused by adverse drug events (ADEs) ranging from 2.4% to 12.0%¹,²,³. Although the safety of a drug must be evaluated in clinical trials prior to its launch, there are potentially many undiscovered side effects due to limitations in terms of both the number of participants and follow-up time. In fact, many drugs have been withdrawn from the market for causing fatal ADEs, such as Vioxx for its doubled risk of causing myocardial infarction⁴ and Cerivastatin for causing fatal rhabdomyolysis⁵. The continued surveillance of drug safety, post marketing, is therefore of utmost importance. Currently, this activity is almost exclusively dependent on spontaneous reports⁶, i.e., voluntary reporting of ADEs by patients or clinicians. Research on using alternative data sources, such as electronic health records (EHRs), are, however, emerging⁷,⁸,⁹. The use of EHRs for pharmacovigilance has several advantages that can complement and compensate for some of the weaknesses of spontaneous reports, such as under-reporting of ADEs and limited access to patient history¹⁰. An important property, and advantage, of EHRs is that they record longitudinal observations of patients and their treatment, including drug use. This opens up great opportunities for analyzing the temporal relationships between ADEs and other clinical events, which, in turn, can aid in the detection of ADEs.

In comparison to spontaneous reports, the reporting of ADEs in EHRs is better regulated, yet the problem of under-reporting remains¹¹. To improve the reporting of ADEs, given the large amounts of data archived in EHR systems, manually screening all records to identify ADEs that were not reported is obviously very costly and even unrealistic. Instead, systems that employ machine learning methods are more efficient – and perhaps not less effective – by learning patterns from patients who have ADEs encoded in their medical history and detecting patients that have suffered an ADE, yet in whose records it has not, but should have, been encoded. Employing such systems could result in improved reporting of ADEs and, in the long run, potentially reduce the unsafe use of drugs.

There are two parallel approaches for such systems to facilitate the reporting of ADEs: (1) to prospectively predict whether there are ADEs to report in new health records; and (2) to retrospectively detect ADEs that are missing in earlier health records. Both approaches require the underlying models to include information from patients’ medical history, given that the first approach needs to use patient history to train the predictive models. The problem thus boils down to making meaningful use of patient history information in EHRs in order to detect ADEs. This problem has previously been studied from different perspectives, such as how best to represent clinical events¹²,¹³. In this study, we focus on the temporal aspect, i.e., which time periods should be considered when collecting clinical events for predicting ADEs. This aspect is expected to have a high impact on the resulting predictive performance, as the probability of an adverse event of a drug occurring at a certain point in time is highly dependent on when the drug was taken, e.g., an ADE will typically not occur for the first time years after a drug has been taken. On the one hand,
considering longer time periods may lead to the ADE signal drowning in the volumes of collected information, while, on the other hand, shorter time periods may exclude information that is crucial for making accurate predictions.

In previous research, a few studies have touched upon or analyzed temporal relationships in EHRs for detecting ADEs. Norén et al. monitor clinical events that have occurred before and after a drug is prescribed to look for abnormalities as signals indicating ADEs\textsuperscript{14}. Their study is, however, in a different genre as it does not involve a learning procedure. At the same time, Jin et al. transform the temporal problem onto a cross-section problem by defining hazard period, effect period and reference period after a drug is prescribed\textsuperscript{15}. In a recent study, Eriksson et al. extract clinical notes and structured prescription data between drug introduction and discontinuation and then filter ADEs based on time stamp inconsistency between structured data and notes\textsuperscript{16}. However, all of these studies mainly look at the temporality on a rather detailed level, such as between specific events or predefined periods; how to handle the temporality in EHRs in general still remains unanswered. In this study, the scope is set to detecting ADEs that should be reported in EHRs by employing machine learning methods to learn patterns using clinical events, and hence the essence of the problem is how to model the events that occurred at different time points in the patient history. To the best of our knowledge, this is the first study to explore ways of handling the temporality of patient history prior to an ADE. The aim of this study is therefore two-fold: (1) to investigate various ways of handling the temporality of clinical events; (2) to explore the importance of clinical events at different time points for detecting ADEs.

Materials and Methods

In this section, we first introduce the strategies of handling temporality that are proposed in this study. Then, we describe a series of empirical experiments that are conducted to evaluate the proposed strategies, starting by describing the data source and subsequently the details of each experiment.

- Strategies to handle temporality

Three strategies – bag of events, bag of binned events, and bag of weighted events – to handle the temporality of clinical events in patients’ medical history for the detection of ADEs were proposed and evaluated in this study. In each of the strategies, only clinical events that occurred prior to the target ADE were included. A toy example of handling the temporality of one drug, one diagnosis and one clinical measurement in a time period of three days is demonstrated in Figure 1.

Here, let $x$ denote each unique event in the whole EHR database, $d$ denote the number of days prior to the occurrence of the target ADE, $d \leq D$, and $n_d$ denote the number of times $x$ occurred in day $d$.

- **Bag of Events (BE)** This strategy counts the number of occurrences of event $x$ within $D$ days. In this case, the value of feature $x$ is $\sum_{d=1}^{D} n_d$.

- **Bag of Binned Events (BBE)** This strategy counts the number of occurrences of event $x$ in each day within $D$ and represents $x$ as $x_1, x_2, ..., x_D$. The value of the corresponding feature here is $n_1, n_2, ..., n_D$.

- **Bag of Weighted Events (BWE)** This strategy assigns different weights to event $x$ that occurred at different days $d$ and takes into account the weights when counting the number of occurrences of $x$. The proportional weights are assigned according to the time distance between the event and the target ADE: events that occurred further in time from the target ADE receive proportionally less weight. In this case, the time distance between $x$ at day $d$ and the target ADE is $d$, and hence the value of feature $x$ is $\sum_{d=1}^{D} (n_d / d)$.

- Data source

In this study, data was extracted from the Stockholm EPR Corpus\textsuperscript{17}, which contains health records of around 700,000 (anonymized) patients over a two-year period (2009-2010), obtained from Karolinska University Hospital in Stockholm, Sweden. This database contains large amounts of diagnosis information, drug administrations, clinical measurements and clinical notes in free-text. In this study, we used only the structured data, i.e., diagnoses, drugs and clinical
measurements, as features. In the Stockholm EPR Corpus, diagnoses are encoded by the International Statistical Classification of Diseases and Related Health Problems, 10th Edition (ICD-10) and drugs are encoded by the Anatomical Therapeutic Chemical Classification System (ATC). This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm) with permission number 2012/834-31/5.

The targeted use case in this study is to detect patients who should, but do not, have a specific ADE reported in their health records by retrospectively analyzing clinical events in their medical history. Among the diagnosis codes that indicate ADEs, 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) were considered in this study. We selected the 14 most frequent A.1 and A.2 codes in the Stockholm EPR Corpus, and thus created 14 datasets, where the existence of the diagnosis code indicating a particular ADE was used as the class label in each dataset. The classification task is hence binary: positive or negative towards a specific ADE. Both inpatients and outpatients are included in this study, where the positive examples are patients whom have been assigned an ADE-specific diagnosis code and the negative examples are patients whom have been assigned a similar diagnosis code to the diagnosis code indicating ADE. Here, similarity is defined as two codes sharing the same first three levels of the ICD-10 concept hierarchy. For instance, if the positive examples are patients with diagnosis code G44.4 (drug-induced headache), the negative examples are patients with any diagnosis code starting with G44 (other headache syndromes), but not G44.4. Table 1 lists the basic descriptions of each dataset, including the diagnosis code that indicates the corresponding ADE (dataset name), the description of this code and the number of positive and negative examples.

Around 10,000 unique ICD-10 diagnosis codes, 1,500 unique ATC codes and 730 unique clinical measurements exist in the Stockholm EPR Corpus. Since most clinical events only occurred to a small group of patients, the datasets where clinical events are used to describe each patient are consequently both high-dimensional and very sparse, i.e., for a single observation (or example) the vast majority of the features have a value of 0. The employed algorithm for generating predictive models, i.e., the random forest algorithm, is rather efficient in handling high-dimensional data, as only a small random sample of the available features is considered when determining the best way of partitioning the training examples during tree growth. However, for highly sparse data, it is not unlikely that all sampled features are uninformative, i.e., lead to no separation. Unless specifically handled, the tree growth will terminate prematurely and lead to an overall low predictive performance. Rather than employing any of the more sophisticated approaches to handle this, we here employed the quite straightforward approach of removing features that are more sparse than 99%, i.e., the ones for which non-zero values were observed in fewer than 1% of the patients; for the datasets with fewer than one hundred observations, features with only one non-zero value were also removed. The motivation behind this is simply that for features with very few non-zero values, the impact of applying different strategies to handle temporality will be negligible, even though some of these features might be valuable indicators.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Corresponding diagnosis code description</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>D642</td>
<td>Secondary sideroblastic anemia due to drugs and toxins</td>
<td>113</td>
<td>4234</td>
</tr>
<tr>
<td>G240</td>
<td>Drug-induced dystonia</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>G444</td>
<td>Drug-induced headache, not elsewhere classified</td>
<td>31</td>
<td>1102</td>
</tr>
<tr>
<td>G620</td>
<td>Drug-induced polyneuropathy</td>
<td>19</td>
<td>367</td>
</tr>
<tr>
<td>I952</td>
<td>Hypotension due to drugs</td>
<td>38</td>
<td>480</td>
</tr>
<tr>
<td>L270</td>
<td>Generalized skin eruption due to drugs and medicaments</td>
<td>174</td>
<td>291</td>
</tr>
<tr>
<td>L271</td>
<td>Localized skin eruption due to drugs and medicaments</td>
<td>58</td>
<td>407</td>
</tr>
<tr>
<td>O355</td>
<td>Maternal care for (suspected) damage to fetus by drugs</td>
<td>334</td>
<td>373</td>
</tr>
<tr>
<td>T782</td>
<td>Adverse effects: anaphylactic shock, unspecified</td>
<td>136</td>
<td>1467</td>
</tr>
<tr>
<td>T783</td>
<td>Adverse effects: angioneurotic oedema</td>
<td>147</td>
<td>1448</td>
</tr>
<tr>
<td>T784</td>
<td>Adverse effects: allergy, unspecified</td>
<td>984</td>
<td>612</td>
</tr>
<tr>
<td>T808</td>
<td>Other complications following infusion, transfusion and therapeutic injection</td>
<td>353</td>
<td>59</td>
</tr>
<tr>
<td>T886</td>
<td>Anaphylactic shock due to correct drug or medicament properly administered</td>
<td>53</td>
<td>607</td>
</tr>
<tr>
<td>T887</td>
<td>Unspecified adverse effect of drug or medicament</td>
<td>472</td>
<td>277</td>
</tr>
</tbody>
</table>

In this study, we defined the following 12 thresholds from patient history: 1, 2, 3, 4, 5, 6, 7, 14, 21, 30, 60, 90 days before the target ADE, where, for each threshold \( n \), clinical events that occurred \( n \) days before the target ADE are studied. The choice was made with the assumption that more clinical events would have occurred in the days closer to the ADE and the difference between days closer to the ADE has a higher impact on detecting ADEs, especially on the ones that are dose-independent. We therefore studied clinical events that occurred in each day of the first week, followed by every week in the first month, and finally every month up to 3 months (90 days) before the occurrence of the target ADE. In the bag of binned events strategy, clinical events are binned within a time threshold. For example, all events that occurred one day before ADE are binned in day 1, and all events that occurred from 60 days to 90 days before the ADE are binned in day 90. Here, the same event that occurred at different times is treated as different events, the number of features in bag of binned events is hence much higher than the other two strategies. To reduce the number of potential models, we do not distinguish between the events that occurred in different days within, e.g., the third month before the ADE. The average number of features and sparsity over the 14 datasets, before and after removing the extremely sparse features at each time threshold, are illustrated in Figure 2.

**Figure 2:** Average number of features and sparsity before and after removing extremely sparse features

- **Experiments**

In this study, a series of experiments was conducted to evaluate the strategies to handle temporality of clinical events in patient history and explore the impact of using different lengths of patient history on detecting ADEs. In the first experiment, the three strategies – bag of events, bag of binned events and bag of weighted events – were compared by employing the random forest\textsuperscript{19} learning algorithm to features (clinical events) generated by each of them respectively.
When more than two competing models are compared, the Friedman test was employed for statistical testing of the null hypothesis that all models perform equally, followed by a post-hoc test using the Bergman-Hommel procedure. The choice of random forest was made due to its reputation of achieving high accuracy, its ability to handle high-dimensional data, as well as the possibility of obtaining estimates of variable importance. The algorithm constructs an ensemble of decision trees, which together vote for what class label to assign to an example that is 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 sampled at each node when building the tree, in both cases to increase diversity among the trees. When the number of trees in the forest increases, 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. Although this condition can only be guaranteed in theory, the algorithm has often been shown in practice to result in state-of-the-art predictive performance. In this study, we used random forest with 500 trees. In all experiments, models were built and evaluated using stratified 5-fold cross validation with two iterations, where the original class distribution was retained in each fold.

The considered performance metrics are accuracy, area under the ROC curve (AUC), precision, recall, F-score and area under the precision-recall curve (AUPRC). Accuracy corresponds to the percentage of correctly classified instances. AUC 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 instance ahead of a negative one. 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 gold 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 means that the algorithm is able to detect most true ADEs. F-score describes the harmony between precision and recall by calculating \( \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \). Only both high precision and high recall can yield a high F-score. AUPRC depicts the probability that precision is higher than recall for each recall threshold. It is considered to be a more careful measurement compared to AUC, since high AUC often leads to high AUPRC, but not necessarily the other way around.

In a subsequent set of experiments, variable importance generated with random forest using bag of binned events was analyzed in order to obtain a deeper understanding of in which time period clinical events are more informative. Variable importance can be estimated in different ways. In this study, Gini importance was chosen as the variable importance metric, where high Gini importance means that a variable plays a greater role in splitting the data into the defined classes. A Gini importance of zero indicates that a variable is considered useless or is never selected to build any tree. Here, we ranked clinical events from different days according to their Gini importance, and then aggregated all events from the same time threshold by averaging their ranks. Given that each time threshold now has an average rank, we calculated the relative importance among the 12 time thresholds: \( D = \{1, 2, 3, 4, 5, 6, 7, 14, 21, 30, 60, 90\} \); for each dataset, the average rank for each time threshold \( d \) is \( r_d, r_d \in R \), and then the relative importance \( r_i \) for time threshold \( d \) is \( r_i = (\max(R) - r_d)/(\max(R) - \min(R)) \), \( r_i \in (0, 1) \). For example, if the relative importance of day 5 is 0 and of day 30 is 0.12, it means that day 30 is 12% more important than day 5.

**Results**

Here, we first report the results on comparing the proposed strategies for handling temporality: bag of events, bag of binned events and bag of weighted events, and then on the impact of clinical events from different days of patient history for each specific ADE.

- **Comparing strategies to handle temporality**

  The predictive performance of random forest with each strategy, in terms of accuracy, AUC, AUPRC, precision, recall and F-score, is presented in Figure 3. We can see that in general, bag of weighted events (in green) outperforms the other two and the predictive performance improves overall with increasing number of days of patient history.

  Friedman tests were conducted to compare the three strategies at each time threshold over the 14 datasets (p-values are shown in Table 2). P-values are not available for time threshold 1 day since the three strategies are identical for
clinical events that occur one day before the ADE. From this table we can see that the p-values generally decrease with more days of patient history, and the null hypothesis that the three strategies perform equally well is rejected with most evaluation metrics towards the end, i.e., using 90 days of patient history. This indicates that when we include more days of patient history in our predictive models, the impact of how temporality is handled increases. Note that although there seems to be a performance drop from 60 days to 90 days in Figure 3, if we take a closer look at the performance on each dataset (e.g., Figure 4 for bag of weighted events), we can see that for most datasets, the predictive performance improves from 60 days to 90 days, except for a big drop on two (G620 and L271), which explains most of the overall performance drop observed in Figure 3.

Given these results, a post-hoc analysis comparing the strategies pairwise was carried out with 90 days as threshold; the results are shown in Table 3. The left half of the table shows the average rank – the lower the rank, the better the performance – among the three strategies, and we can see that bag of weighted events is consistently better than the other two. The pairwise p-values in the right half of the table indicate that bag of weighted events is significantly better than bag of events for AUPRC, precision and F-score; bag of weighted events is significantly better than bag of binned events for AUC and AUPRC; and bag of binned events is significantly better than bag of events for AUC.

- Impact of temporality on specific adverse drug events

The predictive performance of the best observed strategy – bag of weighted events – for each specific ADE is presented in Figure 4. As described in the previous section, the predictive performance overall increases on a small scale towards
Table 3: Average ranks and p-values from the post-hoc analysis of Friedman tests with 90 days as threshold

<table>
<thead>
<tr>
<th>Metric</th>
<th>BE</th>
<th>BBE</th>
<th>BWE</th>
<th>BBE vs. BWE</th>
<th>BE vs. BWE</th>
<th>BBE vs. BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>2.14</td>
<td>2.29</td>
<td>1.57</td>
<td>0.176</td>
<td>0.176</td>
<td>0.705</td>
</tr>
<tr>
<td>AUC</td>
<td>1.96</td>
<td>2.71</td>
<td>1.32</td>
<td>0.0007</td>
<td>0.089</td>
<td>0.047</td>
</tr>
<tr>
<td>AUPRC</td>
<td>2.36</td>
<td>2.43</td>
<td>1.21</td>
<td>0.0004</td>
<td>0.004</td>
<td>0.85</td>
</tr>
<tr>
<td>Precision</td>
<td>2.57</td>
<td>1.89</td>
<td>1.54</td>
<td>0.34</td>
<td>0.018</td>
<td>0.07</td>
</tr>
<tr>
<td>Recall</td>
<td>2.28</td>
<td>1.96</td>
<td>1.75</td>
<td>0.57</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>F-score</td>
<td>2.5</td>
<td>1.96</td>
<td>1.54</td>
<td>0.26</td>
<td>0.032</td>
<td>0.156</td>
</tr>
</tbody>
</table>

using more days of patient history, which is especially clear for D642, T808 and T887. Even with an obvious drop at the end on e.g. G620, the predictive performance with 90 days is still higher than it is with 1 day. Apparently, the choice of threshold has different degrees of impact on different ADEs.

Figure 4: Predictive performance of random forest with bag of weighted events for each ADE

To obtain a deeper understanding of the contribution of clinical events from different days of patient history to the predictive performance, variable importance analysis was conducted using bag of binned events, where they are treated as different variables. Figure 5 shows the relative importance of clinical events occurring in different days of patient history for detecting each specific ADE. For most ADEs, it is clear that clinical events occurring in the late stage of patient history, e.g., just one day before the ADE, is of highest importance; for some ADEs, such as D642 and G620, clinical events occurring in a much earlier stage of patient history are relatively more important.

Discussion

For the task of detecting ADEs using clinical events in EHRs, the predictive performance with the three strategies differs from each other within a fairly small range, even though there are significant differences observed. This can be explained by the difficulty of the task itself. Here, in each dataset, the positive examples are patients who have a specific ADE, while the negative examples are patients who have a disease that is in the same family as the ADE (they share the first three levels of ICD-10). In many cases, these patients are very similar to each other in terms of having similar drugs and clinical measurements; therefore it is difficult for the learning algorithm to distinguish between them. Hence it is difficult to observe big changes in predictive performance with different strategies since they are all bounded with the similar clinical events these patients share.
Based on our results, the random forest built with bag of events has the worst predictive performance, followed by bag of binned events, and both of them are, in most cases, outperformed by the random forest built with bag of weighted events. This finding is consistent with the granularity of each of the strategies: the model with bag of events completely ignores temporality, and hence is the crudest model; although the model with bag of binned events treats events from different time periods as different features, such a representation also results in increased dimensionality and sparsity that have a negative impact on the predictive performance; and, finally, the model with bag of weighted events tackled the problems with the other two strategies – temporality, dimensionality and sparsity – by aggregating the events from different time periods according to their time distance from the target ADE.

The predictive performance and variable importance for each specific ADE to some extent reflect the different nature of each ADE. Adverse drug events are typically divided into two types: dose-dependent or not. The former is related to the accumulation of toxics from drugs or medications; therefore, when predicting ADEs of this type, we would intuitively expect that using events from longer patient history contributes to improved predictive performance compared to using only events from the most recent days. This assumption is supported by the results on predicting D642 – drug induced anemia, where the predictive performance improves monotonically with the use of more days of patient history (see Figure 4 – D642) and the clinical events from seven days prior to the ADE are much more important compared to one day (see Figure 5 – D642). Another example is G620 – drug induced polyneuropathy, which is an ADE known to be caused by drugs used in the treatment of very severe or chronic diseases like cancer or tuberculosis. Such patients are normally exposed for a long period until the side effect is observed. Here, again, a similar performance trend is observed to support our hypothesis (see Figure 4 – G620 and Figure 5 - G620). On the other hand, the second type of ADE is related mainly to idiosyncratic or immunological nature, such as an allergic reaction, which indicates that most likely, with a very marginal assumption, they are more or less instant side effects from taking drugs or medications. The results on predicting T886 – anaphylactic shock due to correct drug or medication properly administered is a good example that agrees with this assumption. The predictive performance almost remains the same as using events from only one day before the ADE (see Figure 4 – T886) and the events from one day before are absolutely the most important ones (see Figure 5 – T886).

However, the assumptions that are made above are not always reflected by our results in this study, which can be explained by several reasons. First, only clinical events from the structured EHRs are used as features here, while ADEs are often described in the notes rather than reported as diagnoses (given that ADEs are heavily under-reported as we described in the introduction). Especially for patients who suffer from severe diseases, ADEs become less
important to report. Therefore, using only the structured clinical events can hardly capture a holistic and precise picture for detecting ADEs. Second, in the real clinical setting, where the EHR data is collected, there is a lack of controls on when clinical events, especially diagnosis codes, are reported in the database. Sometimes events that occurred weeks ago are reported together with the ones that occurred one day ago when the patients are discharged. This, unfortunately, results in noise in the training data and also inaccurate weights based on the time distance from the target ADE. Third, diagnosis codes, i.e., the ICD-10 codes, are themselves heterogeneous. For example, T782 - T784 cover not only ADEs but also adverse reactions to substances that are not drugs.

One limitation of this study concerns the fact that the events are represented as a bag in all of the strategies, which neglects the temporality between the events in the same day or period. This has a potential impact on the predictive performance, especially for predicting ADEs that are not dose-dependent, since the temporal relationship between events within a short period is important to capture when the ADEs are mostly instant effects. To take this into account, techniques that enable mining event sequences would be useful. In addition, the use of ICD-10 codes to select patients with and without ADEs should be proceeded with caution due to the fact that they are coded also for billing purposes and other reasons. For future work, it would also be interesting to see the impact of the proposed strategies on the task of distinguishing patients who have ADEs from patients who do not, i.e., the negative examples would be randomly selected from the population. The way of assigning weights in bag of weighted events is fairly crude and harsh; a follow-up study could focus on exploring alternative ways of assigning weights to clinical events that occur at different time points. To improve the predictive performance, including free-text clinical notes in the models and handling the temporality of information in notes are indeed worth investigating.

Conclusions

It is advantageous to use longitudinal data from electronic health records for the detection of adverse drug events. This study proposed and evaluated three strategies to handle temporality of clinical events: drugs, diagnosis and clinical measurements. These strategies differ from each other in how they take into account the clinical events that occur in different days of patient history: bag of events counts the number of times that an event occurred without regard to when it occurred; bag of binned events separates the patient history into predefined bins and then counts the number of times that an event occurred in each bin; and bag of weighted events counts the weighted number of times that an event occurred, where the weights are assigned proportionally according to time distance between the event and the target adverse drug event. Based on our empirical investigation, the bag of weighted events strategy yields the best predictive performance with the considered metrics. Here, we conclude that, in general, using longer patient history leads to improved predictive performance, and the temporality of clinical events matters more when using more days of patient history.

Acknowledgements

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


