2016 IEEE International Conference on Healthcare Informatics

Predicting Adverse Drug Events using Heterogeneous Event Sequences

Isak Karlsson
Dept. Computer and System Sciences
Stockholm University
Stockholm, Sweden

Henrik Boström
Dept. Computer and System Sciences
Stockholm University
Stockholm, Sweden

***Abstract*—Adverse drug events (ADEs) are known to be severely under-reported in electronic health record (EHR) systems. One approach to mitigate this problem is to employ machine learning methods to detect and signal for potentially missing ADEs, with the aim of increasing reporting rates. There are, however, many challenges involved in constructing prediction models for this task, since data present in health care records is heterogeneous, high dimensional, sparse and temporal. Previous approaches typically employ bag-of-items representations of clinical events that are present in a record, ignoring the temporal aspects. In this paper, we study the problem of classifying heterogeneous and multivariate event sequences using a novel algorithm building on the well known concept of ensemble learning. The proposed approach is empirically evaluated using 27 datasets extracted from a real EHR database with different ADEs present. The results indicate that the proposed approach, which explicitly models the temporal nature of clinical data, can be expected to outperform, in terms of the trade-off between precision and recall, models that do no consider the temporal aspects.**

***Keywords*—Adverse drug events, temporal patterns, data series, ensemble methods, random forest**

I. Introduction

Adverse events caused by the intake of medications account for an increasing amount of hospitalizations and deaths worldwide [3], [16], and is a significant burden on the healthcare system [28]. Harmful adverse events caused by drugs are often referred to as adverse drug events (ADEs), and the activities related to the detection, signaling and assessment of said events is referred to as pharmacovigilance. Although benefit-risk analysis of newly developed drugs is already conducted during clinical trials, post-marketing detection and surveillance is necessary in order to detect unanticipated events, e.g., interacting drugs, since clinical trials are normally performed with a limited sample followed up during a limited period of time. As a result, many drugs have been withdrawn from the market due to serious adverse reactions, e.g, Cerivastatin, a drug for to lower cholesterol and prevent cardiovascular diseases, was withdrawn worldwide in 2001 because of causing fatal rhabdomyolysis [12].

During post-marketing surveillance, a vast array of automatic approaches for detecting potential safety hazards of drugs have been investigated, cf. [1], [26], using various data sources, the most prominent of which is disproportionality analysis of spontaneous individual case reports [30] submitted to e.g., the World Health Organization (WHO)[1]. One of the main obstacles with current systems for collecting data regarding adverse event is the fact that serious ADEs are heavily under-reported (while known ADEs are over-reported) [14] by both clinicians, in the case of EHRs, and by patients, in the case of individual case reports. Studies indicate that as few as $\sim 10\%$ of all serious ADEs are reported [31]. To improve the reporting rate, researchers [36], [37] have investigated systems for automatically detecting possible ADEs from electronic health records (EHRs), which have emerged as a valuable and rich source of data when monitoring the safety of drugs, avoiding several of the limitations present for case reports. For example, EHRs typically contain longitudinal observational data of large samples of patients, including demographic information, medical history, drug consumption with exposure time and dose information, clinical measurements, including lab results and drug concentrations, and clinical narratives evolving over time [9]. Traditionally, various rule based methods have been investigated, e.g., [2], [17]. Recently, however, instead of approaches that require hand-crafted rules, data-driven machine learning methods, which support exploration and discovery of statistical patterns in large databases, have been extensively investigated [18], [19], [22].

Although the rich information stored in EHRs open up for possibilities of signaling ADEs in order to improve reporting rate and patient care using statistical pattern mining methods, there exists several issues with regards to the representation of heterogeneous and time evolving variables, such as clinical measurements, e.g., blood pressure and heart rate or other measurements taken at the local clinic or external laboratories, drug prescriptions and medical conditions. First, since most ADEs are rare and the number of reported cases few, datasets extracted from EHRs for predicting adverse drug events typically contain a small number of examples with thousands of features (corresponding to e.g., drugs, diagnoses and measurements), which in most cases are only recorded for a small fraction of the patients, making such datasets highly *sparse*. Second, since clinical measurements are normally recorded more than once during the medical history of a patient, it is unclear how to handle the temporal nature of such data. One possible solution to address the high-dimensionality is to limit the analysis to a small sample of clinical measurements with high prevalence, which however may severely limit the predictive power. Similarly, the problem of handling temporality may be addressed by only considering a short time-window in the medical history

[1] http://www.who.int

978-1-5090-6117-4/16 $31.00 © 2016 IEEE
DOI 10.1109/ICHI.2016.64

of the patients, which again may result in models with low predictive performance, as such an approach will not exploit one of the most important aspects of EHRs, namely the ability to consider the full medical history of a patient. One way of addressing this, is to construct hand-crafted rules and heuristics, which can be used to derive new features [8]. Another aggregation-based alternative, evaluated by Zhao et al. [36], is to represent clinical measurements using various summary statistics, such as measurement averages (mean and mode for numerical and categorical variables, respectively), the number of measurements and the existence of measurements. The main conclusion in that study is that simply counting the number of measurements during the medical history leads to the highest predictive performance of the considered alternatives [36]. One drawback of such representations is, again, that they only consider a summarized snapshot of the medical history. Under such situations, where multiple variables evolve over time, sequence classification [33] has emerged as a potential approach to handling time evolving events (data series), e.g., to classify electrocardiograms (ECGs) [20], [32], to provide early diagnostics [13] and classifying biological sequences [11].

For numerical sequences, e.g., blood pressure or temperature, the *shapelet* [34] has been introduced as an important primitive for sequence classification. A shapelet is a phase-independent, i.e., location invariant, sub-sequence of a longer time-series used as a local primitive for classification. In this setting, the idea is to find the closest matching position within each time-series to a shapelet and use this distance as a discriminatory feature. For sequence classification, several approaches have been proposed, e.g., shapelet-based decision trees [34], logical combinations of shapelets [24] and shapelet-based transformations [15], [23]. To improve predictive performance, and to support multivariate sequences, the generation of forests of randomized shapelet trees (RSF) [21] has been introduced as a less computationally costly alternative.

Although a random shapelet forest can handle multivariate numerical sequences [21], medical data, as pointed out earlier, often contain disparate types of clinical measurements with high missing rates, i.e., the measurements are recorded only for a small fraction of the patients. In this study, two extensions to the random shapelet forest, which we call the random pattern forest (`RPF`), is explored to address these limitations, i.e., to support heterogeneous and multivariate data series where some sequences (measurements) can be missing from the medical history of a patient. Hence, the purpose of this study is two fold. First, we explore the impact of temporality when predicting adverse drug events extracted from a real database of electronic health records. Second, we empirically evaluate the proposed extensions, in terms of the trade-off between precision and recall (F_1-score) and the accuracy of the predicted probabilities (*Brier score*), for both the proposed method, which models temporality, and for a baseline method based on sparse and aggregated features, which does not.

The remainder of this paper is organised as follows: in Section II, we present background definitions and formulate the problem of classification of heterogeneous event sequences. In Section II-A, we describe the proposed method and in section III, we present an empirical evaluation using 27 datasets extracted from a real EHR system. Finally, in section IV, we summarize the main conclusions and suggest directions for future work.

II. Background

The problem studied in this paper is classification (signaling) of adverse drug events, and our focus is on providing two extensions to the random shapelet forest to handle multivariate and heterogeneous data series. More concretely, the task is, given a set of medical records represented as d-dimensional time evolving data series, where each dimension describe a particular aspect of a patient record, e.g., blood pressure or a sequence of prescribed drugs, we want to infer a classification model that is able to correctly predict the presence of an ADE for a previously unseen medical record. Formally, a *d-dimensional* heterogeneous data series $\mathcal{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_d\}$ is a sequence of d disparate attributes, such that $\mathbf{T}_{ki} \in \mathbb{D}_k$, $\forall k \in \{1, \ldots, d\}$, where $\mathbf{T}_k = \{T_{k,1}, \ldots, T_{k,m_k}\}$ and $\mathbb{D}_k$ is the (numerical or categorical) domain of the k:th attribute. Furthermore, we assume that we have a collection of n multivariate data series $\mathcal{X} = \{(\mathcal{T}_1, y_1), \ldots, (\mathcal{T}_n, y_n)\}$ which defines a *training set*, where each data series is labeled with a label $y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is a finite set of class labels, here representing the presence or absence of a particular ADE.

A. Random Pattern Forest

As briefly introduced above, shapelets are local, phase independent sub-sequences of univariate data series, i.e., shapelets are 1-dimensional numerical subsequences. Originally, shapelets were introduced as local primitives for classification, by considering the closest matching position within each data series to a particular shapelet and use this similarity as a discriminatory feature [34]. The first shapelet classifier [34] embedded the extraction algorithm in a decision tree achieving competitive classification accuracy. To improve the classification accuracy of shapelet trees and reduce learning time, an algorithm for generating ensembles of randomized shapelet trees was presented in [21]. Building on these ideas, we here generalize the algorithm to support any numerical or categorical pattern, given a well defined distance measure.

Ensemble methods rely on the combined voting of several, relatively weak, models which all are assumed to perform better than random guessing and to make somewhat independent errors [6]. In the random pattern forest, randomization in the generation of weak models is introduced both in the selection of training instances and in the selection of patterns to use at each node. The first is performed by employing bagging [5], i.e., randomly selecting n instances with replacement from the original n instances, duplicating some and excluding others. The process results in two disjoint subsets, where each tree is built using the larger in-bag instances and an unbiased estimate of the running performance is given by the out-of-bag instances [6]. The second randomization works by evaluating only a small random sample of patterns at each node [21].

The random pattern forest construction algorithm consists of two parts: ensemble creation and tree generation (see Algorithm 1). The algorithm requires a set of training instances $\mathbf{E} = \{(\mathcal{T}_1, y_1), \ldots, (\mathcal{T}_n, y_n)\}$ where each instance e_i is described by a d-dimensional data series $\mathcal{T}_i \in \mathcal{X}$ and a class label $y_i \in \mathcal{Y}$. Additionally, the algorithm requires two parameters: the number of trees to generate (p) and the number of patterns, e.g., shapelets, to examine at each split (r). In the ensemble part, p pattern trees are constructed from an in-bag sample drawn with replacement from $\mathbf{E}$. The pattern tree construction

Algorithm 1 The random pattern tree algorithm.

procedure PATTERNTREE($\mathbf{E}, r$)
  if $\mathbf{E}$ is pure or other stopping criteria is met **then**
    return the most frequent class label in $\mathbf{E}$
  end if
  for $i \leftarrow 1, n$ **do**
    $\mathbf{p} \leftarrow$ SAMPLEPATTERN($\mathbf{E}$)
    $\mathbf{d} \leftarrow$ COMPUTEDISTANCE($\mathbf{p}$, $\mathbf{E}$)
    compute the information gain of splitting on $\mathbf{p}$
  end for
  $(\mathbf{p}_{best}, \tau_{best}) \leftarrow$ best pattern and distance threshold
  $PT \leftarrow$ create a node that test $\mathbf{p}_{best}$
  $\mathbf{E}_l \leftarrow$ instances with distance to $\mathbf{p}_{best} \leq \tau_{best}$
  $\mathbf{E}_r \leftarrow$ instance with distance to $\mathbf{p}_{best} > \tau_{best}$
  for both l and r as v **do**
    $PT_v \leftarrow$ PATTERNTREE($\mathbf{E}_v$, r)
    Attach PT_v to the corresponding branch of PT
  end for
end procedure

algorithm, which can be run in parallel for each tree to be generated, starts by selecting r random patterns and, using the pair-wise distances between patterns and data series, computes an impurity measure[2] and selects the pattern that reduces error the most if split upon, cf., [21]. The data is subsequently partitioned into two subsets according to the selected pattern and distance threshold, i.e., one subset for those instances with a distance less than the threshold and one for those with a distance greater than the threshold. The tree generation continues recursively to build sub-trees until the resulting nodes are pure, i.e., containing instances of only one class. For example, we obtain the univariate random shapelet algorithm tree algorithm [21] if, in Algorithm 1, SAMPLEPATTERN returns a random shapelet and COMPUTEDISTANCE computes the euclidean sub-sequence distance. To support multiple dimensions, the sampled pattern also indicates which dimension it is sampled from. More specifically, patterns extracted from the k:th dimension is only compared to the k:th data series dimension, e.g.,, the distance between patterns extracted from patients blood pressure is only compared other patients blood pressure. Finally, to handle missing attributes, the maximum (observed) distance is assumed between a pattern and a series that is missing.

B. Pattern extraction and distance measures

In this study, one numerical and two categorical distance measures are evaluated together with three pattern sampling strategies, out of which two are the same for all distance measures and one is only defined for categorical data series. The three proposed pattern sampling strategies, i.e., SAMPLEPATTERN, are:

- (1) *Random* $1d$*-data series:* This strategy samples a data series according to a uniform distribution, i.e., $i \in \mathcal{U}(1, n)$ and $k \in \mathcal{U}(1, d)$, resulting in a single data series dimension $\mathbf{T}_k^i$. This strategy is used for both numerical and categorical data series.
- (2) *Random* $1d$*-data series subsequence:* A data series subsequence of the k^{th} dimension of a data series, $\mathcal{T}$, is a sequence of l contiguous elements of $\mathbf{T}_k$, denoted as $\mathbf{T}_k^{s:s+l-1} = \{T_{k,s}, \ldots, T_{k,s+l-1}\}$, where s is the starting position and l is its length. This strategy samples a data series subsequence from a random data series, $\mathbf{T}_k^i$, by uniformly selecting a length $l \in \mathcal{U}(1, |\mathbf{T}_k^i|)$ and a start position $s \in \mathcal{U}(1, |\mathbf{T}_k^i| - l)$. This strategy is used for both categorical and numerical data series.
- (3) *Random singleton:* From a randomly sampled categorical data series, $\mathbf{T}_k^i$, this strategy sample a single element, $v_k = T_{kj}^i$ according to a uniform distribution.

The distance measure, i.e., the function COMPUTEDISTANCE, is selected differently depending on the type of data series and sampling strategy. For categorical data series, sampled using strategy (1) and (2), we consider the edit distance (denoted as ED) [35], which is defined as

$$d_{(\mathbf{T}_k, \mathbf{T}'_k)}(m, m') = \begin{cases} \max(i, j) \text{ if } \min(i, j) = 0 \\ \min \begin{cases} d(i-1, j) + 1 \\ d(i, j-1) + 1 \\ d(i-1, j-1) + 1_{(T_{ki} \neq T_{kj})} \end{cases} \end{cases} \quad (1)$$

where $m = |\mathbf{T}_k|$ and $m' = |\mathbf{T}'_k|$ and $1_{(T_{ki} \neq T_{kj})}$ is the indicator function. This distance measure is selected to investigate the temporal impact of categorical data. For numerical data series, sampled using strategy (1) and (2), we consider the minimum Euclidean distance (denoted as l^2-norm) between the two (of different length) sequences, i.e.:

$$d_{(\mathbf{S}_k, \mathbf{T}'_k)} = \min_{s=1}^{m-l+1} \{\text{norm}^2(\mathbf{S}_k, \mathbf{T}_k'^{s:s+l-1})\} \quad (2)$$

where $\text{norm}^2 = \|a - b\|^2$, l is $|S|$, m is $|\mathbf{T}_k|$ and S is the shorter of the two data series. Finally, for strategy (3) we consider the zero/one distance (denoted as 0/1), where the maximum distance is assigned if the sampled element is not in the compared data series and 0 otherwise. Mathematically:

$$d_{(e_k, \mathbf{T}_k)} = \begin{cases} 1 & \text{if } e \in \mathbf{T}_k \\ 0 \end{cases} \quad (3)$$

This distance measure does not consider temporality, and is only valid for categorical data.

III. EXPERIMENTS

A. Data source

This paper investigates the possibility to utilize temporal patterns when predicting adverse drug events from EHRs. The investigation is carried out using 27 clinical datasets (see Table I), where the task is to detect care episodes where a particular ADE related diagnosis code[3] should be assigned. The

[2] In this study the information gain is used [27], [34].

[3] International Statistical Classification of Diseases and Related Health Problems, 10th Edition (ICD-10).

TABLE I: Description of the datasets. Each dataset is extracted from the Stockholm EPR corpus.

		Care episodes (n)		**Predictors** (d)		**Unique items** (F)	
Dataset	Code description	#	Frac. of ADE	#	Sparsity	#	Sparsity
D64.2	Secondary sideroblastic anemia due to drugs and toxins	713	0.499	402	0.958	1451	0.983
E27.3	Drug-induced adrenocortical insufficiency	258	0.465	300	0.970	901	0.986
F11.0	Mental and behavioural disorders (MBOs) due to use of opioids: acute intoxication	486	0.422	345	0.971	1190	0.989
F11.2	MBDs due to use of opioids: dependence syndrome	1053	0.466	396	0.977	1780	0.993
F13.0	MBDs due to use of sedatives or hypnotics: acute intoxication	771	0.361	349	0.980	1320	0.993
F13.2	MBOs due to use of sedatives or hypnotics: dependence syndrome	268	0.493	209	0.942	870	0.981
F15.0	MBDs due to use of other stimulants, including caffeine: acute intoxication	238	0.378	199	0.965	562	0.985
F15.1	MBOs due to use of other stimulants, including caffeine: harmful use	208	0.481	195	0.944	646	0.978
F15.2	MHOs due to use of other stimulants, including caffeine: dependence syndrome	656	0.489	343	0.970	1344	0.990
F19.0	MHOs due to multiple drug use: acute intoxication	658	0.404	283	0.976	1137	0.993
F19.2	MBDs due to multiple drug use: dependence syndrome	1167	0.492	420	0.978	2008	0.994
F19.9	MBDs due to multiple drug use: unspecified mental and behavioural disorder	265	0.487	233	0.963	858	0.986
G24.0	Drug Induced Dystonia	187	0.337	220	0.971	623	0.986
G44.4	Drug-induced headache, not elsewhere classified	569	0.469	336	0.984	1079	0.993
G62.0	Drug-Induced Polyneuropathy	347	0.478	286	0.971	907	0.988
I42.7	Cardiomyopathy Due To Drug And External Agent	122	0.467	186	0.945	506	0.973
I95.2	Hypotension Due To Drugs	407	0.484	324	0.962	1139	0.986
L27.0	Generalized skin eruption due to drugs and medicaments	1802	0.478	627	0.981	2746	0.994
L27.1	Localized skin eruption due to drugs and medicaments	614	0.477	343	0.974	1420	0.991
O35.5	Maternal care for (suspected) damage to fetus by drugs	2192	0.500	401	0.990	2007	0.996
T59.9	Toxic effect of unspecified gases, fumes and vapors	397	0.343	229	0.972	820	0.990
T78.2	Adverse effects: anaphylactic shock, unspecified	559	0.453	348	0.984	1061	0.993
T78.3	Adverse effects: angioneurotic oedema	2660	0.431	503	0.989	2225	0.997
T78.4	Adverse effects: allergy, unspecified	12884	0.430	790	0.994	4185	0.999
T80.8	Other complications following infusion, transfusion and therapeutic injection	1554	0.499	452	0.974	2103	0.992
T88.6	Anaphylactic shock due to correct drug or medicament properly administered	371	0.488	294	0.964	1114	0.987
T88.7	Unspecified adverse effect of drug or medicament	4305	0.457	801	0.989	3696	0.997

ADE-related diagnosis codes used as the target variable were selected based on prevalence (more than 100 care episodes) and classified as indicating ADEs in a previous study [29]. The 27 datasets are extracted from the Stockholm EPR Corpus[4], which contains medical records for over $1,000,000$ patients, admitted to one of 512 clinical units within Stockholm City Council during seven consecutive years (2007-2014) [9]. In the database, health care episodes are described by both clinical narratives and structured data regarding prescribed drugs[5], diagnoses and clinical measurements. In this study, each dataset is comprised of prescribed drugs, assigned diagnosis codes, clinical measurements, e.g., blood pressure, creatinine levels and albumin. In Table I, the number of care episodes and the relative class frequencies are listed for each dataset, together with the number of predictors, i.e., data series of ATC and ICD codes and measurements, and their missing rate, i.e., fraction of measurements that exist per care episode. In the last columns of the table, the number of unique items[6] and their sparsities are listed.

The empirical evaluation primarily concerns the evaluation of two representations of this data, i) a bag-of-items representation, where each care episode is represented as a frequency distribution over items, and ii) as multivariate data series, where each care episode is represented by numerical and categorical event sequences. The first representation ignores temporal dependencies and is utilized by the Random Forest [6] algorithm, while the second representation exploits the temporal aspects using the novel random pattern algorithm introduced here. The different approaches are compared in terms of the trade-off between precision and recall (sensitivity) and the accuracy of the predicted probabilities.

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

[5]Anatomical Therapeutic Chemical classification system (ATC).

[6]One item is a particular ATC code, ICD code or measurement.

TABLE II: The configurations of distance measures and sampling strategies used in the experiments. For all approaches, 10, 25, 50 and 100 ensemble members were evaluated.

Approach	COMPUTEDISTANCE	SAMPLEPATTERN	Note
RPF-ed$^{(1)}$	ED, l^2-norm	(1), (1)	$r = d * 0.01$
RPF-ed$^{(2)}$	ED, l^2-norm	(2), (2)	$r = d * 0.01$
RPF$^{(3,\ 1)}$	0/1, l^2-norm	(3), (2)	$r = d * 0.01$
RPF$^{(3,\ 2)}$	0/1, l^2-norm	(3), (1)	$r = d * 0.01$
RF	–	–	$mtry = \sqrt{F}$

B. Experimental setup

To evaluate the predictive performance of machine learning models, an independent test set should be employed. If there are plenty of data available, the simplest and most common approach is to split the data into two halves: one for training the model and the other for evaluating the predictive performance of the model. If data is scarce, which is true for many medical domains, cross-validation is another viable approach for estimating the predictive performance. Cross-validation works by partitioning the data into k disjoint subsets, which are iteratively used to train a model on $k-1$ partitions and evaluate the model using 1 partition averaging the performance over k train and test iterations. A usual choice for k is 10, which is also employed in this study. Since both the baseline, i.e., Random Forest, classifier and the novel, Random Pattern Forest, classifier are stochastic, the acquired measurements are averaged over 5 rounds of 10-fold-cross-validation to minimize the variability.

For the experiment, we consider four configurations of the proposed pattern forest. As seen in Table II, all approaches employ Euclidean distance (Eq. 2) for numerical data. For categorical data, on the other hand, either the edit distance (ED) (Eq. 1) or the 0/1-distance is used. For the first approach, RPF-ed$^{(1)}$, full data series are sampled according to approach (1) for both numeric and categorical data series and compared

using ED; for the second approach, $\texttt{RPF-ed}^{(2)}$ subsequences are sampled according to approach (2) and compared using ED. For the third approach, $\texttt{RPF}^{(3,\,1)}$, numerical data series are sampled according to approach (1) and singleton values (approach (3)) are sampled for categorical data series and compared using the 0/1-distance (Eq. 3). Finally, the fourth approach ($\texttt{RPF}^{(3,\,2)}$) samples categorical singleton values, similarly to the third approach, but numerical data series are instead sampled according to approach (2). For the baseline approach, we consider the number of times each event is recorded as feature since this was found to be the best performing feature aggregation approach [36].

Both the the baseline and the proposed algorithm have a few parameters to be set. Firstly, to investigate the effect of ensemble size, i.e., the number of trees, an increasing number of trees are generated for each approach. The considered ensemble sizes are 10, 25, 50, and 100, where the largest ensemble size has been suggested as an appropriately sized ensemble [25]. Secondly, for the baseline approach the number of features sampled at each node is set to the square-root of the number of possible features and for the novel approach, the number of sampled pattern, at each node, is set to 1% of the available dimensions (see Table II).

The most commonly employed metric for evaluating the predictive performance of machine learning models is the percentage of correctly predicted examples, i.e., the accuracy. However, in many cases, e.g., when the the cost of true positives and false negatives are unknown, the precision ($p = tp/(tp + fp)$ and recall (sensitivity) ($r = tp/(tp + fn)$) allow for a more meaningful comparison. The F_1-score captures the trade-off between these and is defined as $F_1 = (2pr)/(p + r)$. Finally, when evaluating different alternatives, e.g., whether a patient has an ADE or not, the accuracy of the predicted probabilities is often as important as the overall accuracy. One commonly employed metric when assessing the accuracy of probabilities is the (mean) squared error, also known as Brier score [7]. Let c_i be the true class vector of an instance and $c_{ij} = 1$ if $y_i = c_{ij}$ and 0 otherwise. Furthermore, let p_i be a vector of probabilities assigned by the classifier to each class for the i:th instance, then the square error is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\|c_i - p_i\|^2. \tag{4}$$

By decomposing the mean squared error into two terms, the bias and the variance, it is possible to investigate what explains the differences in predicted performance. Given the true class vector c of an instance and the mean probability vector $\hat{p}_\mu$, the mean square error of a random ensemble can be defined as (cf. [4]):

$$\frac{1}{p}\sum_{t=1}^{p}\|\hat{p}_t - c\|^2 = \|c - \hat{p}_\mu\|^2 + \frac{1}{p}\sum_{t=1}^{p}\|\hat{p}_t - \hat{p}_\mu\|^2 \tag{5}$$

where p is the number of members in the ensemble. In the decomposition, the first term on the right hand side is the Brier score and the second term the variance. Hence, in an ensemble, one wants to minimize the squared error while increasing the variance.

To detect if there are any significant differences in predictive performance, as measured by F_1-score, brier score and variance, between the different approaches, the widely accepted non-parametric Friedman test based on ranks over the datasets [10] is employed. To detect differences between pairs of approaches, a Nemenyi post-hoc test is performed [10].

C. Empirical Results

The mean rank for the baseline approach (RF) and temporal approaches are shown in Table III. To provide an overview of the results, the average performance is given for each measurement in Fig. 1 and 2. Note, however, that the statistical test for determining the significant differences, are conducted based on the ranks among the individual datasets and approaches.

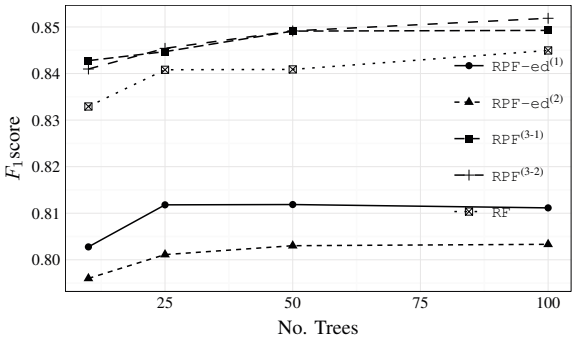

Fig. 1: Average F_1-score for all approaches and ensemble size.

Examining the mean F_1-score (Fig. 1), one can see that the approaches utilizing temporal dependencies for categorical data, i.e., RPF-ed, clearly perform the worst. The best performing approaches consider the temporal aspect for numerical features only, i.e., RPF. For all ensemble sizes, a Nemenyi test reveals that both RPF-ed approaches are significantly outperformed ($p < 0.01$) by all alternative approaches. No significant differences can, however, be detected between RF and RPF for any ensemble size.

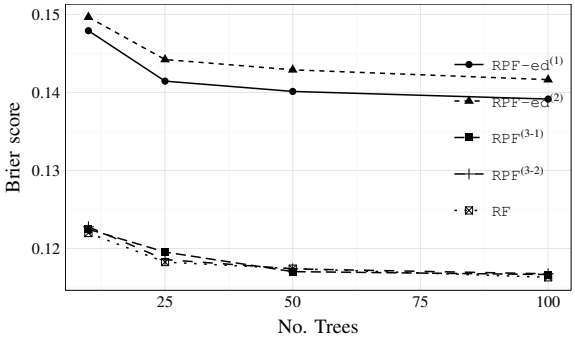

Fig. 2: Average accuracy of the predicted probabilities (brier score) for all approaches and each forest size.

Considering Brier score (Fig. 2), we can see that the predicted probabilities are similarly accurate for both RF and

TABLE III: Mean rank for the considered approaches, with regard to F_1-score, Brier score and variance for all ensemble sizes.

	F_1-score				Brier				Variance			
Approach	10	25	50	100	10	25	50	100	10	25	50	100
RPF-ed$^{(1)}$	3.79	3.87	3.74	4.12	4.08	3.93	4.08	4.08	2.00	2.11	2.07	2.11
RPF-ed$^{(2)}$	4.40	4.16	4.37	4.57	4.15	4.23	4.12	4.23	**1.92**	**1.66**	**1.74**	**1.66**
RPF$^{(3\text{-}1)}$	2.33	**2.12**	2.25	2.46	2.41	2.67	2.41	2.38	3.18	3.40	3.40	3.37
RPF$^{(3\text{-}2)}$	**2.09**	2.31	**1.98**	**1.74**	2.41	2.23	2.38	2.41	3.29	3.22	3.14	3.25
RF	2.37	2.51	2.64	2.09	**1.97**	**1.97**	**2.04**	**1.93**	4.59	4.59	4.62	4.59

RPF, but significantly less accurate for RPF-ed. Indeed, a Friedman test shows that the observed differences deviate significantly ($p < 0.01$) from what can be expected under the null-hypothesis of no difference. More specifically, as seen in Table III, while both RPF and RF significantly ($p < 0.01$) outperform RPF-ed for all ensemble sizes, there is no significant difference between the approaches that consider the temporality for numerical data series and the baseline approach that does not take temporal patterns into consideration.

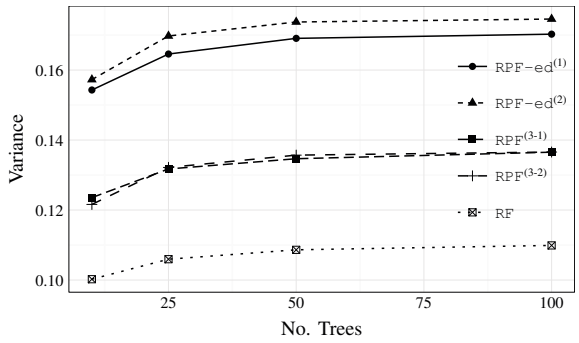

Fig. 3: Average variance of the predicted probabilities for all approaches and for each ensemble size.

Finally, in Fig. 3, the variance part of the mean square error of predicted probabilities in Eq.5 is shown. A Friedman test reveals that all deviations are significant with a *p*-value < 0.01. A Nemenyi test shows that, for all forest sizes, the differences when comparing RPF-ed to RPF and RF are significant (the first has higher variance than the latter). Conclusively, any benefit of the introduced novel approach which considers temporal numerical data series, compared to the baseline approach with regard to F_1-score, can be attributed to the fact that the novel approach increases the variance of the ensemble without simultaneously increasing the bias too much compared to the baseline approach. From a qualitative perspective, one can see that the novel approach outperform the baseline for predicting conditions that can be detected from sudden changes in various medical measurements. For example, the trade-off between precision and sensitivity for predicting drug induced adrenocortical insufficiency (E27.3) is improved by 5% points compared to the baseline. For other reactions, where sudden changes in measurements cannot be detected, e.g., generalized skin eruption due to drugs and medicaments (L27.0), the predictive performance is on the other hand decreased by 3% points. Hence, while the general predictive performance is not significantly increased by the proposed method, for certain types of ADEs it can improve the detection rates.

IV. Concluding Remarks

In this paper, a novel ensemble approach, which extracts and compares temporal patterns in a decision forest, is empirically evaluated for cases where the task is to predict whether or not a health care episode should be assigned an ADE related diagnosis code, based on data extracted from an EHR database.

The empirical results show that models that take into consideration temporal patterns for numerical data series, perform better than the baseline approach, which ignores temporality. However, considering the temporality of categorical data series does not seem improve predictive performance, as measured by F_1-score. One reason for the poor result when utilizing categorical temporal patterns, might be the fact that the order in which drugs have been prescribed and diagnoses assigned does not reveal any additional information. For numerical measurements, the effectiveness of treatment and also possible negative effects of drugs might on the contrary be revealed by the evolution of such measurements.

Although the result does not clearly show that the proposed approaches for considering temporality when predicting adverse drug events significantly outperform the baseline approach, which ignores the temporality, when measuring the F_1-score and Brier score, the study highlights the importance of investigating new ways of handling the temporality of clinical measurements. As shown in the empirical investigation, by decomposing the mean square error of the ensembles into two parts, the Brier score and the variance, the differences in predictive performance of the different methods can be explained by the fact that while ignoring temporality results in pattern trees that are more accurate on average, considering temporality results in more variability in the individual predictions leading to an overall increased performance. One suggestion for future work is to investigate how the increased variability can be utilized to improve the predictive performance.

One limitation of the current study, possibly affecting the results, is the fact that while most patients have been prescribed drugs and assigned diagnoses, most clinical measurements are missing for a large fraction of patients, which could explain the slight difference between the baseline approach and the novel approach. Hence, for future studies, it would be important to investigate the importance of both patterns and dimensions when making predictions using the novel approach. This is also important for increasing the acceptance among practitioners. Another important direction for future studies could be to consider alternative distance measures and pattern sampling strategies that take into consideration domain specific knowledge. One such approach, could for example be to represent health care episodes as graphs of related events, allowing for more elaborate representations.

ACKNOWLEDGMENT

This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by Swedish Foundation for Strategic Research under grant IIS11-0053.

REFERENCES

[1] JS Almenoff, EN Pattishall, TG Gibbs, W DuMouchel, SJW Evans, and N Yuen. Novel statistical tools for monitoring the safety of marketed drugs. *Clinical Pharmacology & Therapeutics*, 82(2):157–166, 2007.

[2] Andrew Bate, Marie Lindquist, IR Edwards, Sten Olsson, Roland Orre, Anders Lansner, and R Melhado De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, 54(4):315–321, 1998.

[3] HJM Beijer and CJ De Blaey. Hospitalisations caused by adverse drug reactions (adr): a meta-analysis of observational studies. *Pharmacy World and Science*, 24(2):46–54, 2002.

[4] Henrik Boström. Forests of probability estimation trees. *International journal of pattern recognition and artificial intelligence*, 26(02):1251001, 2012.

[5] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[6] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[8] Emmanuel Chazard, Grégoire Ficheur, Stéphanie Bernonville, Michel Luyckx, and Régis Beuscart. Data mining to generate adverse drug events detection rules. *Information Technology in Biomedicine, IEEE Transactions on*, 15(6):823–830, 2011.

[9] Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. Stockholm epr corpus: A clinical database used to improve health care. In *Swedish Language Technology Conference*, volume 18. Citeseer, 2012.

[10] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[11] Mukund Deshpande and George Karypis. Evaluation of techniques for classifying biological sequences. In *Advances in Knowledge Discovery and Data Mining*, pages 417–431. Springer, 2002.

[12] Curt D Furberg and Bertram Pitt. Withdrawal of cerivastatin from the world market. *Curr Control Trials Cardiovasc Med*, 2(5):205–207, 2001.

[13] Mohamed F Ghalwash, Vladan Radosavljevic, and Zoran Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 201–210. IEEE, 2013.

[14] Lorna Hazell and Saad AW Shakir. Under-reporting of adverse drug reactions. *Drug Safety*, 29(5):385–396, 2006.

[15] Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. Classification of time series by shapelet transformation. *Data Mining and Know. Discovery*, 28(4), 2014.

[16] RL Howard, AJ Avery, S Slavenburg, S Royal, G Pipe, P Lucassen, and M Pirmohamed. Which drugs cause preventable admissions to hospital? a systematic review. *British journal of clinical pharmacology*, 63(2):136–147, 2007.

[17] Ashish K Jha, Gilad J Kuperman, Eve Rittenberg, Jonathan M Teich, and David W Bates. Identifying hospital admissions due to adverse drug events using a computer-based monitor. *Pharmacoepidemiology and drug safety*, 10(2):113–119, 2001.

[18] Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):56, 2015.

[19] Isak Karlsson and Henrik Bostrom. Handling sparsity with random forests when predicting adverse drug events from electronic health records. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*, pages 17–22. IEEE, 2014.

[20] Isak Karlsson, Panagiotis Papapetrou, and Lars Asker. Multi-channel ecg classification using forests of randomized shapelet trees. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, page 43. ACM, 2015.

[21] Isak Karlsson, Panagotis Papapetrou, and Henrik Boström. Forests of randomized shapelet trees. In *Statistical Learning and Data Sciences*, pages 126–136. Springer, 2015.

[22] Isak Karlsson, Jing Zhao, Lars Asker, and Henrik Boström. Predicting adverse drug events by analyzing electronic patient records. In *Artificial Intelligence in Medicine*, pages 125–129. Springer Berlin Heidelberg, 2013.

[23] Jason Lines, Luke M Davis, Jon Hills, and Anthony Bagnall. A shapelet transform for time series classification. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 289–297. ACM, 2012.

[24] Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical-shapelets: an expressive primitive for time series classification. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1154–1162. ACM, 2011.

[25] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition*, pages 154–168. Springer, 2012.

[26] Antoine Pariente, Fleur Gregoire, Annie Fourrier-Reglat, Françoise Haramburu, and Nicholas Moore. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases the notoriety bias. *Drug safety*, 30(10):891–898, 2007.

[27] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 1993.

[28] Sebastian Schneeweiss, Joerg Hasford, Martin Göttler, Annemarie Hoffmann, Ann-Kathrin Riethling, and Jerry 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*, 58(4):285–291, 2002.

[29] Jürgen Stausberg and Joerg 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*, 11(1):1, 2011.

[30] Ayako Suzuki, Raul J Andrade, Einar Bjornsson, M Isabel Lucena, William M Lee, Nancy A Yuen, Christine M Hunt, and James W Freston. Drugs associated with hepatotoxicity and their reporting frequency of liver adverse events in vigibase™. *Drug safety*, 33(6):503–522, 2010.

[31] Meredith Wadman. News feature: strong medicine. *Nature medicine*, 11(5):465–466, 2005.

[32] Li Wei and Eamonn Keogh. Semi-supervised time series classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 748–753. ACM, 2006.

[33] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010.

[34] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proc. of the 15th ACM SIGKDD*. ACM, 2009.

[35] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1091–1095, 2007.

[36] Jing Zhao, Aron Henriksson, Lars Asker, and Henrik Bostrom. Detecting adverse drug events with multiple representations of clinical measurements. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 536–543. IEEE, 2014.

[37] Jing Zhao, Aron Henriksson, Lars Asker, and Henrik Boström. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC medical informatics and decision making*, 15(Suppl 4):S1, 2015.