Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records

Reinier Kop[a,*], Mark Hoogendoorn[a], Annette ten Teije[a], Frederike L. Büchner[b], Pauline Slottje[c], Leon M.G. Moons[d], Mattijs E. Numans[b,c,e]

[a] VU University Amsterdam, Department of Computer Science, Amsterdam, The Netherlands
[b] Leiden University Medical Center, Department of Public Health and Primary Care, Leiden, The Netherlands
[c] VU University Medical Center, Academic Network of General Practice, Department of General Practice and Elderly Care Medicine, Amsterdam, The Netherlands
[d] Utrecht University Medical Center, Department of Gastroenterology and Hepatology, Utrecht, The Netherlands
[e] Utrecht University Medical Center, Julius Center of Health Sciences and Primary Care, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form
14 June 2016
Accepted 20 June 2016

Keywords:
Colorectal cancer
Data mining
Data processing
Electronic medical records
Machine learning

ABSTRACT

Over the past years, research utilizing routine care data extracted from Electronic Medical Records (EMRs) has increased tremendously. Yet there are no straightforward, standardized strategies for pre-processing these data. We propose a dedicated medical pre-processing pipeline aimed at taking on many problems and opportunities contained within EMR data, such as their temporal, inaccurate and incomplete nature. The pipeline is demonstrated on a dataset of routinely recorded data in general practice EMRs of over 260,000 patients, in which the occurrence of colorectal cancer (CRC) is predicted using various machine learning techniques (i.e., CART, LR, RF) and subsets of the data. CRC is a common type of cancer, of which early detection has proven to be important yet challenging.

The results are threefold. First, the predictive models generated using our pipeline reconfirmed known predictors and identified new, medically plausible, predictors derived from the cardiovascular and metabolic disease domain, validating the pipeline's effectiveness. Second, the difference between the best model generated by the data-driven subset (AUC 0.891) and the best model generated by the current state of the art hypothesis-driven subset (AUC 0.864) is statistically significant at the 95% confidence interval level. Third, the pipeline itself is highly generic and independent of the specific disease targeted and the EMR used. In conclusion, the application of established machine learning techniques in combination with the proposed pipeline on EMRs has great potential to enhance disease prediction, and hence early detection and intervention in medical practice.

1. Introduction

Predictive models for diseases can greatly contribute to the domain of health. They can help to identify high risk groups and diseases in an early stage, be a facilitator for proactive care, and assist in selecting the most effective treatment. In the current era, more and more medical data are stored electronically allowing more accurate predictive models to be generated. Amongst others, these data include detailed medication prescriptions, laboratory results, coded and free-text consultation visits. Traditional, more hypothesis-driven, predictive model approaches from the medical and epidemiological domain are still applicable and valuable. However, they no longer necessarily lead to the best possible predictive models as they do not fully utilize the wealth of information contained within the EMRs.

This is where the domain of machine learning comes into play. Algorithms originating from that field are well-suited to maximize usage of the variety of information stored within EMRs and work in a data-driven way contrary to the aforementioned hypothesis-driven approaches. However, even for these sophisticated machine learning approaches, extracting useful predictors from EMRs is not a trivial task due to the very nature of the data. First of all, the data is of a highly temporal nature, whereby consecutive events are stored and linked to individual patient records such as consultations, prescribed medication, referrals and lab results. In addition, the data is typically incomplete caused by (1) the decision of the patient whether or not to present complaints, (2) a physician's observational competence, (3) a physician's registration routines and (4) the type of EMR system being used. Finally, certain values stored in the system cannot easily be interpreted if not seen in a

[*] Correspondence to: De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.
E-mail address: r.kop@vu.nl (R. Kop).

http://dx.doi.org/10.1016/j.compbiomed.2016.06.019
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

context (e.g., a laboratory measurement value lacking a unit or reference values). All of these characteristics make it very difficult to apply off-the-shelf machine learning algorithms.

The aim of this work is to develop a dedicated pre-processing pipeline (consisting of a number of components) that is able to address all the aforementioned issues independent of the EMR used. The pipeline combines several pre-processing algorithms. In addition, we develop a novel approach to enrich the EMR data based on knowledge stored in medical ontologies. To evaluate the pipeline, we study the performance of various machine learning techniques in conjunction with our pipeline by applying them to the case of predicting colorectal cancer (CRC) based on a general practitioner (GP) EMR dataset with over 260,000 subjects. The task of predicting CRC occurrence is extremely relevant as it has the third highest incidence among all cancer types worldwide [29]. Early detection is important for survival and quality of life. However, it is also challenging as the symptoms are mostly non-specific and show considerable overlap with benign disease.

Various approaches have been proposed to cope with some of the aforementioned issues. For example, Patnaik et al. [31] and Batal et al. [3] developed approaches that exploit the temporal dimension. However, none of these approaches have tried to cover the entire range of problems in full at once. To the best of our knowledge, the techniques have not been applied to GP data in a very extensive way to show the benefit of the approaches in that context.

This paper is organized as follows. In Section 2 an overview of related work is presented. Section 3 presents the pipeline, including its individual components. The experimental setup to evaluate the approach is presented in Section 4, followed by the results in Section 5. Section 6 is a discussion, and we conclude in Section 7.

2. Related work

EMR data consist of various data types, and often contain missing or wrong data [15]. Processing the data so that samples of uniform length are created is an important first step if traditional predictive modeling techniques are to be used. Though EMR data are timestamped in nearly all cases, it has been shown that clinical prediction tasks have reasonable performance when ignoring the data's temporality (e.g., [8,20,12]), at times even outperforming hypothesis-driven approaches (e.g., [30]). An important limitation in some of these studies is the data-driven approach cannot be considered purely data-driven. Rather, the available data is often already tailored towards the disease or disorder under investigation ([20], and to some extent, [8,30]). For example, Kurt et al. [20] investigate the presence of coronary artery disease using features known to be good predictors for the disease. Such comparisons, though useful to showcase the potential of data-driven prediction in EMRs, are less interesting because new potential predictors will not be found. The present work allows for this as the available primary care data contains information not specific to any particular disease.

To further improve prediction quality, it is potentially useful to apply temporal pattern mining on EMR datasets. Temporal patterns are implicitly contained in EMR data, but to allow traditional prediction methods to use these patterns as features, patterns must be mined in advance. Temporal pattern mining can be viewed as a subtype of association learning [1] in the sense that we are interested in (temporal) relations between events rather than items (see e.g., [34,16,7,26,3]). Batal et al. [3] build progressively larger temporal patterns using a modified version of Agrawal and Srikant's apriori algorithm. However, this algorithm is applied on a dataset spanning days as opposed to months. Kop et al. [18] validate the effectiveness of their method for longer periods of time by reporting an increase in performance when applying their algorithm on EMR data spanning months.

Another way to find additional features is to look at the vast amount of semantic data available in ontologies on the web (e.g., SNOMED, the Systematized Nomenclature of Medicine). Melton et al. [24] use SNOMED to find patient similarity using semantic links between concepts. In La-Ongsri and Roddick [28], EMR concepts are mapped to ontology concepts with the goal to allow multiple levels of abstraction for efficient database usage. This work explores the usage of ontologies in another way: to generate additional features used in a prediction task.

This work is a continuation of previous work in which we explored different subsets of the current data. In Hoogendoorn et al. [12] the potential of EMR data was validated using established machine learning techniques and simple (non-temporal) pre-processing, already resulting in predictive performances regarding CRC better than solely relying on age/gender. This is often difficult within EMR data, as age and gender are known to be among the most obvious predictors in clinical prediction tasks. In Kop et al. [18], this research was expanded on by applying temporal pattern mining, further improving performance. The work currently described builds upon the above by (1) formalizing the entire pipeline, (2) adding lab measurement contextualization, (3) adding semantic enrichment of the data and (4) applying all this on a larger dataset. Furthermore, it attempts to validate the temporal pattern mining work of Batal et al. [3] and reinforces the potential of data-driven research.

Finally, worth mentioning are the analytical platforms Informatics for Integrating Biology and the Bedside (i2b2, [27]) and, by extension, Shared Health Research Information Network (SHRINE, [37]) that allow physicians and researchers to filter and analyze medical data. The important difference is that those platforms are built mostly for human users to better observe and understand their data, whereas our pipeline transforms medical data in order to automatically generate models. In theory, those platforms could be extended to incorporate a pre-processing pipeline such as the one described in this paper. This would allow for large-scale data mining on aggregated medical data sets, which is in line with our research goals.

3. Methodology

In this Section, we introduce our pre-processing pipeline. The code for the pipeline has been made available online.[1] The pipeline is composed of four steps, as shown in Fig. 1. The setup is highly generic and independent of the specific disease targeted and the used EMR. The data present in the EMR includes time stamped events in these categories: consultations, medication, referrals, and values of laboratory measurements. Formally, we specify a set of measurements $a_1, \ldots, a_m$ and a set of patients within our dataset as $p_1, \ldots, p_n$. The domain (i.e., possible values) of a measurement a_i is denoted by A_i. In addition we assume a number of time points $t_{start}, \ldots, t_{end}$ where time is considered in days and the duration is the same for all patients. The value of a measurement i as a specific time point t for patient p is denoted by $a_i(p, t)$. Finally, the type of measurement is specified by means of the type function: $type(a_i)$ which can take the values *consultation*, *medication*, *referral*, and *lab*. Each of the measurements has values set according to some coding scheme to classify the data (e.g., International Classification for Primary Care (ICPC; [4]), for symptoms and diagnosis or Anatomical Therapeutic Chemical (ATC) classification system for

[1] See https://github.com/ReinierKop/EMR-pre-processing-pipeline.

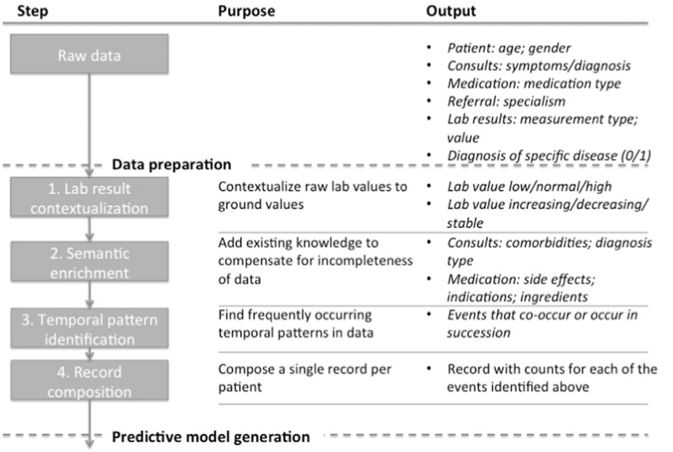
Step
Purpose
Output
Raw data
Patient: age; gender
Consults: symptoms/diagnosis
Medication: medication type
Referral: specialism
Lab results: measurement type; value
Diagnosis of specific disease (0/1)
Data preparation
1. Lab result contextualization
Contextualize raw lab values to ground values
Lab value low/normal/high
Lab value increasing/decreasing/stable
2. Semantic enrichment
Add existing knowledge to compensate for incompleteness of data
Consults: comorbidities; diagnosis type
Medication: side effects; indications; ingredients
3. Temporal pattern identification
Find frequently occurring temporal patterns in data
Events that co-occur or occur in succession
4. Record composition
Compose a single record per patient
Record with counts for each of the events identified above
Predictive model generation

Fig. 1. The pre-processing pipeline to better handle the nature of EMR data while still allowing for deployment of off-the-shelf machine learning techniques. The left column shows the steps involved in the pre-processing, whereas the middle column shows the purpose of each step. The right column shows the output of the step. Italics indicate output in the form of a time-stamped event. Note that the steps, except for the last one, extend the raw dataset and do not replace any previous part of the data.

medication prescription in the case of our CRC study). Such a coding scheme can be advantageous because it provides medical staff with a unified way of entering data in the system, in an attempt to combat inaccuracy and incompleteness. Below, we briefly describe each of the subsequent steps of the pipeline followed by a description of the experimental setup using the pipeline.

3.1. Step 1: Contextualization

The first step comprises the *contextualization of lab results* and aims to provide grounding to values that have been recorded with respect to lab measurements, i.e., it strives to solve the *context* problem identified before. We focus only on the measurements of lab results, i.e., the set $L = \{a_i : type(a_i) = lab\}$. There are two ways to abstract the raw measurement value. The first is *absolute grounding* and is only applicable if reference values $r_{min}(a_i)$ and $r_{max}(a_i)$ are known for that measurement:

$$L_{max} = \{a_i : a_i \in L \wedge r_{min}(a_i) \neq unknown \wedge r_{max}(a_i) \neq unknown\}$$

In other words, we know the allowed minimum and maximum values of a lab measurement. We use a function that changes the domain of the variable to a nominal value: $ground_{abs}: A_i \rightarrow A_{ground\_abs}$ where A_{ground_abs} contains the values low (l), normal (n), or high (h). A new value is assigned as follows:

$$ground_{abs}(a_i(p,t)) = \begin{cases} l, & if \quad a_i(p,t) < r_{min}(a_i) \\ h, & if \quad a_i(p,t) > r_{max}(a_i) \\ n & \quad otherwise \end{cases}$$

The second is *relative grounding* based on pairs of lab measurement of the same patient $(a_i(p, t_1), a_i(p, t_2))$ where $t_1 < t_2$. In the relative grounding, we want to derive whether the values are decreasing, increasing or remain stable when comparing them in pairs. To avoid relatively small fluctuations of values to be considered as increasing or decreasing we define the maximum range of variation of the measurement under consideration as the difference between the lowest and the highest value seen in the database for that measurement:

$$max(a_i) = \max_{\forall p \in \{p_1,\dots,p_n\} \forall t \in \{t_{start},\dots,t_{end}\}} a_i(p,t)$$

$$min(a_i) = \min_{\forall p \in \{p_1,\dots,p_n\} \forall t \in \{t_{start},\dots,t_{end}\}} a_i(p,t)$$

Thus, relative grounding can always be used regardless of the availability of reference values. The $ground_{rel}: A_i \times A_i \rightarrow A_{ground\_abs}$ where A_{ground_abs} are decreasing (d), increasing (i) or stable (s), is calculated as follows:

$$ground_{rel}(a_i(p,t_1), a_i(p,t_2)) = \begin{cases} d, & \text{if } a_i(p,t_2) - a_i(p,t_1) < (\min(a_i) - \max(a_i) \cdot tp) \\ i, & \text{if } a_i(p,t_2) - a_i(p,t_1) < (\max(a_i) - \min(a_i) \cdot tp) \\ s & \text{otherwise} \end{cases}$$

We assign this new value to the attribute value of the second (latest) time point. The parameter *tp* can be used to change the bandwidth of what is considered to be stable.

3.2. Step 2: Semantic Enrichment

The next step in the pipeline involves the *semantic enrichment*. One of the problems with EMR data, as mentioned before, involves the inherent *incompleteness* in the data. One way to tackle this problem is to enrich the dataset by automatically adding domain knowledge using medical ontologies. Essentially, we try to derive (1) additional features that have not been recorded in the database but might be predictive, and (2) try to compensate for data that should be present but might potentially be missing. We chose to access these ontologies using various SPARQL endpoints [33] on the web. In our current pipeline we use the consultation and medication data as they were considered the most logical starting point according to the medical experts, given that lab results are difficult to process and link, and referrals being considered less relevant. With respect to the derivation of additional features we have focused on symptoms associated with a diagnosis (grouping diseases with common symptoms), side effects of medication (allowing for consideration of registered diagnosis that might be a consequence of medication), active ingredients (grouping medication with similar workings). Furthermore, to compensate for missing information we add diseases associated with certain medication (to compensate for diagnoses that have not been registered).

Formally, the semantic enrichment takes a value as input, and outputs the value of a new attribute (referred to as A_{m+1} here) as output: $enrich: A_i \rightarrow A_{m+1}$. Table 1 shows a more detailed overview of the enrichments. A problem is that for most types of enrichment, these endpoints return a too large number of possibilities. For example, the drug N07BB04 can be prescribed for twelve diseases varying from asthma to dementia. This might not be particularly useful, thus this number is reduced by comparing whether the occurrence of the enrichment in the CRC and non-CRC group is significantly different using the two-proportion z test $(\alpha = 0.05)$, and pruning any enrichment that is represented equally in both groups. Significantly different enrichments are added to the dataset with a time stamp equal to its source. Evidently, it is hard to guarantee that even these enrichments are applicable to a particular patient as there is often a wealth of known relationships of which only a subset might be relevant. Our experiments as described later will show whether this is the case. It is important to note that free text in the EMR data has also been shown to provide a form of data enrichment (e.g., [9,35,38]), this is however beyond the scope of our current research which focuses on coded data. As a final note, it is important to realize that this step is meant for recovering randomly missing data (e.g., a physician forgets to include a diagnosis code) and not for recovering

Table 1
Overview of performed enrichment, their sources, purpose, and example outcomes.

Source	Enrichment	Medical ontologies used	Purpose	Example outcome
Consultation	Associated diseases/symptoms	ICPC[a]/SNOMED[b] [36]	Investigate comorbidity	S15 (foreign body in skin): dermatosis
Medication	Associated diseases	DrugBank[c] [17]	Investigate comorbidity	N07BB04: asthma, dementia, malaria,...
Medication	Side effects	DrugBank/SIDER[d] [19]	Investigate symptom causes	S01FA05: heat intolerance, ...
Medication	Active ingredients	DrugBank/SIDER/DailyMed[e]	Investigate ingredient effects	P02CA01: Mebendazole

[a] See http://www.drugbank.ca.
[b] See http://bioportal.bioontology.org/ontologies/SNOMEDCT.
[c] See http://www.drugbank.ca.
[d] See http://sideeffects.embl.de.
[e] See https://dailymed.nlm.nih.gov/dailymed.

structurally missing data. Instead, the assumption is that this structurally missing data is at least partially solvable by incorporating data from various sources, as was done in this study.

3.3. Step 3: Temporal Pattern Identification

The third step in the process addresses the *highly temporal nature* of the EMR data and is referred to as the *temporal pattern identification step*. This step is based on an approach proposed by Batal et al. [3]. During this step, temporal patterns are generated that occur sufficiently frequent based on co-occurrence (c) and succession (s) of events; we do not use Allen's thirteen temporal relations [2] as our data is so incomplete and noisy and these were considered the most relevant (cf. [25]). These relations still allow us to study symptomatic progression, disease comorbidity, and conspicuous lab results. Evidently, this temporal information should already be contained in the dataset in some manner, but this step makes it explicit and accessible for standard learning algorithms. It again adds features to our dataset, but now based on temporal_abstraction: $A_1 \times \ldots \times A_m \rightarrow A_{m+1} \times \ldots \times A_{m+q}$. An example of a simple pattern is (*diagnosis:ICPC_D12 (c) prescription: ATC_A06*), interpreted as the subject has a consult diagnosis of obstipation (D12), as well as received a prescription of anti-constipation drugs (A06) at the same time. It is important to realize that an *(s)* or (*c*) relation is assigned to *each pair* of events in a pattern (cf. [14]). This results in $n(n-1)/2$ assigned relations for a given pattern, where n is the number of events.

The association mining algorithm used to mine such patterns from the data is based on the APRIORI algorithm [1]; the records are first scanned to create frequent *1*-patterns (patterns of size one) f_1. Here, frequent is defined as a minimum percentage for which the pattern occurs across all patients, i.e., the minimum support σ. The frequent patterns are used as input to generate successively larger patterns. In general, the algorithm alternates between two phases:

1. The candidate generation phase uses frequent k-patterns f_k to obtain candidate $k+1$-patterns c_{k+1}. By definition, new candidates are generated by *prepending* p_1 to frequent patterns p where $p_1 \in f_1$ and $p \in f_k$. Then, all possible relational permutations are considered on the newly generated candidates. As an example, prepending *p₁=(ICPC_D01)* to *p=(ICPC_D12 (c) ATC_A06)* requires two new relations to be set, resulting in $|\{c, s\}| \times 2 = 4$ relational permutations, all potentially viable candidates (for numerous optimizations regarding this step, refer to [3]).
2. In the counting phase, the support s is calculated for each member of c_{k+1} by performing a frequency count; each infrequent pattern ($s(p) < \sigma$) is discarded, whereas the frequent patterns ($s(p) \geq \sigma$) become the input for the new candidate generation iteration, i.e., $f_{k+1} = \{p | p \in c_{k+1} \wedge s(p) \geq \sigma\}$. Finally, k is incremented.

The algorithm terminates once a counting phase reports no new frequent patterns ($f_{k+1} = \emptyset$). Larger patterns mined using this algorithm are able to express complex sequences such as “the subject goes to the GP complaining of constipation and heartburn; the subject picks up a prescribed medicine that same day; after these events, the subject returns to the GP with severe stomachache; finally, the GP refers the subject to a gastro-intestinal specialist.” To maximize the likelihood of finding predictive patterns, the algorithm is run on both groups of patients separately (in our case: CRC/non-CRC). The resulting patterns of all sizes for both patient classes are then assembled into a single list. Finally, each resulting pattern is re-interpreted as a binary feature indicating the absence or presence of the pattern in a patient's event history. Due to the explosion of (c) relations when including enriched features (a certain date might have an abundance of registered prescriptions, associated diseases, side effects, and active ingredients which all come from a single event), the algorithm from Batal et al. [3] was adapted to only consider patterns with enrichments from different sources.

3.4. Step 4: Record Composition

To allow application of established machine learning techniques for model generation, all data derived during the previous steps is aggregated into one input vector per patient of uniform length. First, the target variable (in our case absence or presence of CRC) is derived based on the first known registration of the disease in the raw consultation events recorded by the GP. All other raw events in the dataset (i.e., coded consults, prescriptions and referrals) become features for which the number of occurrences of the event for a patient becomes the value. For lab abstraction events, each combination of measurement and abstraction type is aggregated identical to the raw events (note that e.g., *white blood cell count – high* is a different feature than *white blood cell count – normal*). The temporal patterns become binary features indicating the presence or absence of the pattern in a patient, as mentioned. Finally, any general demographics (such as age and gender) are added as features. To improve the scalability, feature selection is applied on the resulting features using the Pearson coefficient [11]. Here, two different feature sets can be used as input for the feature selection: 1) all features from all steps, or 2) only the generated temporal patterns that result from step 3 above. The latter option makes sense as these patterns are essentially a re-interpretation of the regular data from the other steps.

To specify this formally, we assume the set of all attributes that result of the steps expressed before has size s, and the domain of an attribute i is expressed by A_i (as used before). After all transformations, the domains only contain categorical attributes except for the basic demographic attributes (that do not require aggregation). The following features are then distinguished:

$$F = \{ age, gender \} \cup \bigcup_{i=0}^{s} \{ name(a_i)_{v_k} :\ v_k \in A_i \}$$

Note that the values are preceded by the name of the attribute (indicated by "*name*(a_i)"). Furthermore, the value of a (non-demographic) feature for a patient p is specified as follows:

$$value\left(name(a_i)_{v_k}, p\right) = \sum_{t=t_{start}}^{t_{end}} \begin{cases} 1 & if\ \ a_i(t, p) = v_k \\ 0 & otherwise \end{cases}$$

Temporal attributes are binary and not aggregations. In other words, the above value gets truncated to the range $[0, 1]$ for those attributes.

4. Experimental setup

The pipeline described above was applied on the aforementioned GP EMR dataset in order to predict the occurrence of CRC. The dataset is a merger from six anonymized datasets originating from three urban regions in the Netherlands.[2] Each set covers a dedicated GP recording system. The various standards used across the sets were carefully combined with the help of medical experts. Data is available regarding a patient's age, gender, details of GP consultations (> 11 million records in total), drug prescriptions (> 23 million), specialist or additional diagnostic procedure referrals (> 4.4 million), and lab test outcomes (> 22 million), over a period of five years (2007–2011). Note that no data was available regarding dietary habits, nor was free text available for use due to privacy concerns. The lab results table in particular proved challenging to combine, as the measurements were only partially standardized. In the end, 90% of all lab results records were successfully mapped onto each other.

Before feeding this data into the pipeline, we made a sub-selection of suitable patients and corresponding events. First, we selected only those patients aged 30 and up to stay in line with existing literature (cf. [23]). For patients with CRC we selected a period of six months directly prior to the first diagnosis of CRC, with the medically informed expectation that chronic disease and conditions as well as recently developed symptomatology would then be detected as predictors. For patients not diagnosed with CRC we randomly selected a period of six months from the EMR instead. All patients with a too short relevant history (i.e., less than six months) were excluded. The filtered dataset consisted of approximately 263,879 patients with a total of 1292 CRC cases (0.50%), which is in line with both specialist expectation and the reported incidence in the Netherlands.[3] This dataset was then fed to the pipeline, generating abstractions, enrichments, and temporal patterns as described earlier. A transition point *tp* of 0.10 was chosen for the relative grounding in the lab result contextualization step. For the temporal patterns, a minimum support σ of 0.05 was chosen.

The largest processed dataset fed to the algorithms consisted of over 260.000 patients with over 2500 features per patient. The runtime of this was infeasible, especially for the more complex algorithms. Therefore, feature selection using the Pearson correlation coefficient was done. Multiple values for the maximum number of features k were tried, ranging from $k = 50$ to $k = 1000$. The results presented are with $k = 50$, because no significant differences were observed with $k > 50$. Feature selection with $k < 50$ was not experimented with, as ideally, we want to find both previously known CRC predictors as determined by the literature (e.g., [23,40]), about thirty, as well as retaining the opportunity to identify new ones.

After pre-processing, a comparison was made (using the AUC and corresponding 95% confidence intervals, confusion matrices, precision, recall, and F1-score) between three established machine learning techniques that are known to provide insightful results for physicians, i.e., CART, [5], logistic regression (LR), and random forests (RF, [6]). The algorithms are used off-the-shelf using the Python module scikit-learn [32], with parameters chosen as follows, after extensive tuning. For LR, the objective function minimized is the mean squared error. For CART, the Gini impurity measure was used as a splitting criterion; information gain resulted in slightly worse models. A maximum tree depth of 5 was chosen, with a minimum number of 50 samples per leaf node; larger trees and a reduction of the minimum number of samples per leaf node resulted in tremendous overfitting. For RF, the settings as in CART were chosen, with a forest size of 100. The analysis was performed using 5-fold stratified cross validation. The AUCs presented are averages over these folds. Because of the imbalanced nature of the data (0.50% of patients had CRC), all algorithms generated models using a weighting strategy, giving more importance to less frequently occurring classes (CRC in our case), inversely proportional to their occurrence in the data, i.e., $w_{neg} = \frac{\sum_{i=1}^{n} crc}{n}$ and $w_{pos} = 1 - w_{neg}$, where w_{neg} is the weight given to non-CRC cases, w_{pos} is the weight given to CRC cases, *crc* is a binary indicating the presence (1) or absence (0) of CRC in a patient, and n is the total number of patients in the dataset.

Confusion matrices and corresponding precision, recall, and F1-scores are calculated using a predetermined cutoff point, which is essentially the point where the subject's status changes from "unlikely to have the disease" to "likely to have the disease". It has important consequences to set such a cutoff point in a risk model when lives are on the line. Depending on the disease investigated, it might be better to allow for a high false positive rate to detect as many ill patients as possible. In other cases, medical procedures might be too costly or invasive to allow a generous false positive rate. In consultation with medical experts, we chose the cutoff point to generate the confusion matrices and precision, recall and F1-score so that the false positive rate was 0.10.

To test the added value of the pipeline, the algorithms were also applied on a subset of the data only containing available features known a priori to be good predictors for the occurrence of CRC according to the Bristol-Birmingham equation (BBE) as proposed by Marshall et al. [23]. Furthermore, we studied the impact of various segments of the pipeline by leaving out all features of one or more pre-processing steps (i.e., semantic enrichment or temporal pattern identification). Excluding the semantic enrichment when generating temporal patterns will likely change the number of patterns. A final condition we explored was to include only those features generated in the temporal pattern identification step, plus age and gender. Finally, we performed a qualitative analysis of the best performing condition on 100% of the data, i.e., all data was used for generating the final model.

5. Results

With $\sigma = 0.05$, the temporal pattern algorithm generated 125 temporal patterns when including the enrichment step, and 117 when excluding this step. The results are shown in Table 2. The pipeline shows most impact on the AUC when used with LR, specifically on conditions that include temporal pattern identification (step 3). Solely using the temporal patterns provides the best overall performance throughout the entire experiment

[2] Julius General Practitioners' Network, Utrecht; Academic Network of General Practice VU University Medical Center Amsterdam (ANH VUmc); Leiden General Practitioner Registration Network RNUH-LEO, LUMC, Leiden.

[3] See http://www.cijfersoverkanker.nl.

Table 2
AUCs and 95% confidence intervals for the applied algorithms, established using 5-fold cross validation. Conditions are split into three categories: (1) benchmarks (age & gender; BBE); (2) the application of the various steps of the regular pipeline, and (3) like category (2), but only using the resulting temporal patterns and not the other features. The asterisk indicates the condition most similar to the original BBE [23]. The carets indicate the best overall performing conditions.

Algorithm **Condition**	**CART**	**Logistic regression**	**Random forest**
Benchmark			
Age & Gender	0.831 (0.817–0.844)	0.836 (0.823-0.850)	0.834 (0.821–0.848)
Bristol-Birmingham equation	0.851 (0.838–0.864)	*0.864 (0.851–0.877)	0.885 (0.873–0.897)
Regular pipeline			
Steps 1	0.854 (0.841–0.867)	0.871 (0.859–0.884)	0.888 (0.876–0.900)
Steps 1,2	0.854 (0.841–0.867)	0.871 (0.859–0.884)	0.889 (0.877–0.900)
Steps 1,3	0.850 (0.837–0.863)	0.883 (0.871–0.895)	0.881 (0.869–0.893)
Steps 1,2,3	0.849 (0.836–0.863)	0.879 (0.867–0.891)	0.881 (0.868–0.891)
Pipeline with only temporal patterns used for prediction			
Steps 1,3	0.855 (0.842–0.868)	^0.891 (0.879–0.903)	0.882 (0.870–0.894)
Steps 1,2,3	0.855 (0.842–0.868)	^0.891 (0.879–0.903)	0.883 (0.871–0.895)

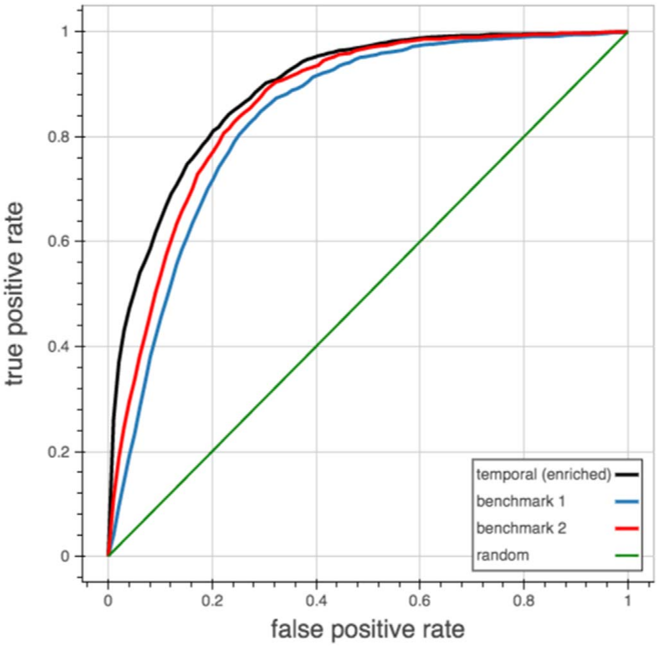

Fig. 2. ROC curves of various generated models. The black line indicates the highest performing model-dataset combination, namely LR applied on temporal pattern data only (with or without enrichment). The red and blue lines are LR applied on benchmarks 1 (Age & Gender) and 2 (BBE), respectively.

Table 3
Confusion matrices for logistic regression models generated using (a) the Age & Gender benchmark data, AUC=0.836 (b) the BBE benchmark data, AUC=0.864, and (c) temporal data only (with or without enrichment), being the highest performing model-dataset combination with AUC=0.891, corresponding to the black line in Fig. 2.

(a)

Actual prediction	1	0
1	588	26,524
0	704	236,063

(b)

Actual prediction	1	0
1	694	26,524
0	598	236,063

(c)

Actual prediction	1	0
1	830	26,524
0	462	236,063

(bottom two conditions, AUC=0.891, precision=0.030, recall=0.642, F1-score=0.058; conditions indicated by carets in the table). Furthermore, the conditions result in a significant improvement compared to both the Age & Gender benchmark (AUC=0.836, precision=0.022, recall=0.455, F1-score=0.041) and the BBE benchmark (AUC=0.864, precision=0.026, recall=0.537, F1-score=0.049). ROC curves corresponding to these AUCs are shown in Fig. 2, and confusion matrices in Table 3. Though the differences between the curves are small, they are still clearly visible. Furthermore, the confusion matrices show a clear improvement in number of identified CRC cases as their corresponding AUCs increase. The latter of the benchmarks, the BBE benchmark, is most similar to the method used in Marshall et al. [23] (i.e., LR performed on BBE predictors only), indicated by an asterisk in the table.

It can be seen that the pipeline is least useful in combination with CART (maximum AUC=0.885, with precision=0.026, recall=0.539, F1-score=0.049). CART is not capable of improving significantly beyond either benchmark and has limited recall and precision. For RF, the level of sophistication of the algorithm results in high base performances (maximum AUC=0.889 with corresponding precision=0.030, recall=0.637, F1-score=0.057), moving past the age/gender benchmark (AUC=0.834, precision=0.022, recall=0.458, F1-score=0.042). However, it performs on par with the BBE benchmark (AUC=0.885, precision=0.029, recall=0.623, F1-score=0.056), meaning the pipeline does not have a clear added value in combination with this algorithm.

The predictors of the best performing model, namely LR with pipeline steps 1,2,3 and 1,3 using temporal patterns only (including those of length one) are listed in.

Table 4 (AUC=0.891 for both conditions). Unfortunately, no semantic enrichment features were found to be predictive, meaning.

Table 4 is applicable to both conditions. It is reassuring to see that the model identifies established CRC alarm symptoms such as anemia or changing bowel habits detected by the use of anti-constipation drugs (indicated by asterisks). Furthermore, hypertension and diabetes are related to the metabolic syndrome, which has been shown to be associated with an increased risk of colorectal cancer [10]; relevant predictors are indicated by carets in.

Table 4
Predictors of the model using solely the temporal patterns (including those of size one) in combination with LR generated using all steps of the pipeline. The importance factor expresses the weight in the model (only those above 0.25 are shown) for that specific pattern. The rightmost columns show the occurrences amongst CRC and non-CRC cases. Predictors prefixed by an asterisk are predictors established in the literature such as the BBE. Those prefixed by carets warrant further medical analysis to establish their relation to CRC. Unmarked predictors have no obvious medical explanation independent of the other predictors. (*s*) represents a succession relationship within a pattern.

Predictor (temporal pattern)	Importance in model	CRC cases	Non-CRC cases
*Drugs for constipation	2.72	336 (26.0%)	7231 (2.8%)
*Iron deficiency anemia	1.88	87 (6.7%)	1080 (0.4%)
*Lipid modifying agents (*s*) *Drugs for constipation	1.77	85 (6.6%)	1098 (0.4%)
*Age	1.62	–	–
*Drugs for acid related disorders (*s*) *Drugs for constipation	1.55	99 (7.7%)	2179 (0.8%)
*Diabetes non-insulin dependent	0.99	161 (12.5%)	8968 (3.4%)
*Abdominal pain/cramps general	0.91	87 (6.7%)	1687 (0.6%)
*Diabetes non-insulin dependent (*s*) *Diabetes non-insulin dependent	0.89	132 (10.2%)	7366 (2.8%)
^Beta blocking agents (*s*) *Drugs for constipation	0.86	68 (5.3%)	1111 (0.4%)
^Hypertension uncomplicated (*s*) ^Hypertension uncomplicated	0.79	168 (13.0%)	9921 (3,8%)
*Agents acting on the renin-angiotensin system (*s*) *Drugs for constipation	0.63	83 (6.4%)	1208 (0.5%)
*Diuretics	0.54	176 (13.6%)	11,497 (4.4%)
Flu vaccination[a]	0.44	228 (17.6%)	19,073 (7.3%)
^Agents acting on the renin-angiotensin system (*s*) *Antithrombotic agents	0.28	97 (7.5%)	5139 (2.0%)
*Abdominal pain localized other	0.28	83 (6.4%)	3056 (1.2%)
General consult (*s*) General consult (*s*) General consult	0.27	75 (5.8%)	3554 (1.4%)
*Agents acting on the renin-angiotensin system (*s*) *Drugs for acid-related disorders	0.26	82 (6.3%)	4071 (1.6%)
^Agents acting on the renin-angiotensin system	0.26	226 (17.5%)	15,824 (6.0%)

[a] Rather than the vaccination as such, this predictor reflects the patients who qualify for influenza vaccination, i.e. those at risk including the elderly and those with pulmonary, cardiovascular and metabolic chronic diseases.

The reported AUCs lie relatively close to one another, often barely (or not even) surpassing the simplest benchmark, age & gender. This is quite common in medical prediction tasks, as age and gender are known to be good predictors for diseases (see e.g., [39] for a breakdown of this regarding CRC). The goal is thus to improve on this benchmark, in which small gains are already of significant added value. Our results show LR and RF accomplish such gains using the pipeline, whereas CART does not. While the best performing condition of LR (AUC=0.891) is not significantly better than the best RF condition (AUC=0.889), the relative gain here is that LR is much more insightful than RF (in our setup, RF contains 100 trees, which is incomprehensible for humans). Such insightfulness is imperative in order to facilitate a physician's trust in the model. To conclude, the pipeline can lift the performance of a simple, yet insightful, classifier (LR) to the level of a sophisticated but less insightful classifier (RF) and even shows a significant increase compared to the BBE benchmark while RF does not.

Unfortunately, the semantic enrichment step did not significantly contribute to the end result. The initial expectation was that (for example) a certain side effect might be prevalent across many different types of drugs, causing it to be discovered as a CRC predictor. However, either the data was not adequate or the enriched concepts were not complete enough to allow for such behavior. Likely contributing to this outcome is the fact that most ontologies do not provide prevalence information on the concepts related to one another (e.g., an extremely rare side effect of a drug such as blindness versus a common side effect such as red eyes), introducing a large number of useless features.

6. Discussion

Although the mapping between the GP recording systems was carefully made, it is inevitable that some concepts cannot be mapped onto one another. This was especially prevalent within the lab measurement table, in which 90% of the concepts were mapped in the end. It is unknown if and how the mapped (and unmapped) concepts influenced results.

A number of design choices have been made with regard to the pipeline, particularly the contextualization step. Firstly, we opted against a higher granularity of the current analyses, or introducing new ones such as the rate of change, since a single lab measurement currently already introduces multiple features (those resulting from absolute and relative grounding). Related to this, if the granularity is set too fine, it could theoretically hamper subsequent temporal pattern generation, as the minimum support for patterns with less frequently occurring lab measurements is less likely to be satisfied by the pattern mining algorithm. Another pragmatic choice we made was to judge all measurements using the same equation, meaning we assumed equal importance across all lab measurements, which might not be realistic. However, it is infeasible to tune the weights of those hundreds of lab measurements in collaboration with medical experts. Finally, we also assumed mutual independence in the case of the lab measurements. Although this is unrealistic as well, implementing such dependence would lead to a combinatorial explosion as well as careful tuning, making it even less feasible than the previous point. The equal importance and mutual independence assumptions may have resulted in subtle changes in more relevant lab measurements to be overlooked.

As mentioned before, the target variable was derived from the raw consultation events. However, since it is plausible the EMRs have missing data, there could also be missing or incorrect data regarding the diagnosis of CRC. Theoretically, the actual diagnosis could have occurred before the recorded date, which would mean values in Table 4 are not CRC predictors, but rather descriptors; they would describe the problems and medication during the course of CRC, rather than what occurred before. Though it is difficult to prevent this entirely, we have examined the resulting predictors in close collaboration with medical experts and are confident the individual predictors more closely resemble consults and prescriptions prior to an initial CRC diagnosis rather than after.

7. Conclusion

The work described here formalizes the use of a pre-processing pipeline applied in the medical domain, which is an attempt to tackle various problems in EMR data. Predictive models are in some cases already applied in practice, but these usually rely on hypothesis-driven approaches (e.g., [23,40]). The pipeline covers incompleteness and inaccuracies (e.g., by contextualizing lab measurements and enriching content), and temporality (by discovering temporal patterns in patients), making it suitable as a data-driven alternative to e.g., the Hippisley-Cox approach. When used on an EMR dataset for predicting

CRC, the pipeline generated features that:

1. Improved some simple insightful algorithms (i.e., LR) to the level of a more sophisticated algorithm (i.e., RF), which works towards improving physicians' trust in predictive models by maintaining insightfulness.
2. Outperformed a pure hypothesis-driven approach (the Bristol-Birmingham equation by [23]) using a data-driven approach, demonstrating the potential of using EMR data for prediction tasks.
3. Rediscovered established CRC predictors, validating the applicability of the generated models in combination with the pipeline and further improving physicians' trust.
4. Found associations between CRC predictors on the one hand, and the metabolic syndrome on the other hand, showing the pipeline's competitiveness compared to recent hypothesis-driven medical research [10].

We plan to further develop the pipeline by adding new data such as clinical free text annotations (see [13]) and dietary habits. Moreover, there is room for improvement in the current steps, such as refining the granularity of lab measurement values in the contextualization step, taking into account their relative importance, and looking at possible mutual interactions between them. Furthermore, the pipeline is in theory applicable for any combination of disease and EMR system, but this needs to be validated. Finally, using the proposed pipeline, pre-selection of patients according to their overall risk of (developing) CRC can take place with data available in EMRs, without additional efforts from patient or physician. As such, considering the pipeline and the generated models for use in combination with national screening programs (e.g., [22] in Finland or [21] in England) is also part of future work.

Author contributions

R.K. has operationalized and implemented the ideas in the machine learning pipeline. M.H. has been the main inventor of the machine learning pipeline and guided operational choices in the pipeline. A.t.T. has contributed ideas on the semantic enrichment step in the pipeline. P.S., F.L.B. and M.E.N. delivered data; P.S., M.E.N. and L.M.G.M. interpreted the data and resulting predictive model, L.M.G.M. and M.E.N. provided medical expert advice, L.M.G.M., M.E.N. and M.H. designed the study, R.K. and M.H. drafted, A.t.T., F.L.B., L.M.G.M., M.E.N., M.H. and P.S. revised the manuscript for important intellectual content, all authors approved the final version and agree to be accountable for all aspects of the work.

Conflict of interest statement

None declared.

Acknowledgments

This project has been partially funded by Dutch Health Care Insurance Company VGZ [Grant no. P6500/D3932]. We thank the GPs from the Julius General Practitioners' Network (JHN), the Academic Network of General Practice VU Medical Center Amsterdam (ANH-VUmc) and the Leiden General Practice Registration Network (RNUH-LEO LUMC) in the Netherlands for sharing anonymized routine general practice care EMR datasets for the purpose of this study. Scamander Solutions provided for support in data management. We thank Tanya Levy (VGZ), Stefan Visser (VGZ), Daoud Urdu (Scamander), Bram Wesselo (Scamander), and Ruud Wijmans (Scamander) for their input and feedback during the research project.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases VLDB, vol. 1215, Sep 12, 1994, pp. 487–499.
[2] J.F. Allen, Towards a general theory of action and time, Artif. Intell. 23 (2) (1984) 123–154.
[3] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, ACM Trans. Intell Syst. Technol. (2013) 4.
[4] Bent Guttorm Bentsen, International classification of primary care, Scand. J. Primary Health Care 4 (1) (1986) 43–50.
[5] L. Breiman, J. Friedman, R. Olshen, C. Stone, D. Steinberg, P. Colla, Cart: Classification and regression trees, Wadsworth, Belmont, CA, 1983.
[6] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[7] S. Concaro, L. Sacchi, C. Cerra, R. Bellazzi, Mining administrative and clinical diabetes data with temporal association rules, InMIE (2009) 574–578.
[8] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.
[9] S. DeLisle, B. South, J.A. Anthony, E. Kalp, A. Gundlapallli, F.C. Curriero, Greg E. Glass, Matthew Samore, T.M. Perl, Combining free text and structured electronic medical record entries to detect acute respiratory infections, PloS one 5 (10) (2010) e13377.
[10] K. Esposito, P. Chiodini, A. Colao, A. Lenzi, D. Giugliano, Metabolic syndrome and risk of cancer: a systematic review and meta-analysis, Diabetes Care 35 (2012) 2402–2411.
[11] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[12] M. Hoogendoorn, L.M. Moons, M.E. Numan,s R.J. Sips, Utilizing data mining for predictive modeling of colorectal cancer using electronic medical records, in: Brain Informatics and Health, Springer International Publishing, 2014, pp. 132–141.
[13] M. Hoogendoorn, P. Szolovits, L.M. Moons, M.E. Numans, Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer, Artif. Intell. Med. (2016).
[14] Frank Höppner, Knowledge discovery from sequential data (Ph.D. Thesis) Diss., Technical University Braunschweig, Germany, 2003.
[15] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[16] H. Jin, J. Chen, H. He, G.J. Williams, C. Kelman, C.M. O'Keefe, Mining unexpected temporal associations: applications in detecting adverse drug reactions, IEEE Trans. Inf. Technol. Biomed. 12 (4) (2008) 488–500.
[17] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, et al., DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, Nucleic Acids Res. 39 (2011) D1035–D1041.
[18] R. Kop, M. Hoogendoorn, L.M. Moons, M.E. Numans, A. ten Teije, On the advantage of using dedicated data mining techniques to predict colorectal cancer, in: Artificial Intelligence in Medicine, Springer International Publishing, 2015, pp. 133–142.
[19] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, P. Bork, A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (2010) 343.
[20] I. Kurt, M. Ture, A.T. Kurum, Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, Expert Syst. Appl. 34 (1) (2008) 366–374.
[21] R.F. Logan, J. Patnick, C. Nickerson, L. Coleman, M.D. Rutter, C. von Wagner, English Bowel Cancer Screening Evaluation C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests, Gut 61 (2012) 1439–1446.
[22] N. Malila, T. Oivanen, O. Malminiemi, M. Hakama, Test, episode, and programme sensitivities of screening for colorectal cancer as a public health policy in Finland: experimental design, BMJ 337 (2008) a2261.
[23] T. Marshall, R. Lancashire, D. Sharp, T.J. Peters, K.K. Cheng, W. Hamilton, The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance, Gut 60 (2011) 1242–1248.
[24] G.B. Melton, S. Parsons, F.P. Morrison, A.S. Rothschild, M. Markatou, G. Hripcsak, Inter-patient distance metrics using SNOMED CT defining relationships, J. Biomed. Inform. 39 (6) (2006) 697–705.
[25] F. Moerchen, Algorithms for time series knowledge mining, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006 ,Aug 20. pp. 668-673.
[26] R. Moskovitch, Y. Shahar, Medical temporal-knowledge discovery via temporal abstraction, in: AMIA, 2009, Nov 14.
[27] S.N. Murphy, M. Mendis, K. Hackett, R. Kuttan, W. Pan, L. Phillips, V. Gainer, D. Berkowicz, J.P. Glaser, I.S. Kohane, H.C. Chueh. Architecture of the open-source

clinical research chart from Informatics for Integrating Biology and the Bedside, in: AMIA, 2007, Oct 11.

[28] S. La-Ongsri, J.F. Roddick, Incorporating ontology-based semantics into conceptual modelling, Inf. Syst. 52 (2015 30) 1–20.

[29] J.B. O'Connell, M.A. Maggard, C.Y. Ko, Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging, J. Natl. Cancer Inst. 96 (2004) 1420–1425.

[30] A. Oztekin, D. Delen, Z.J. Kong, Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology, Int. J. Med. Inform. 78 (12) (2009) e84–e96.

[31] D. Patnaik, P. Butler, N. Ramakrishnan, L. Parida L, B.J. Keller, D.A. Hanauer (Eds), Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[33] E. Prud'Hommeaux, A. Seaborne, SPARQL query language for RDF, W3C Recommendation, 2008, p. 15.

[34] L. Sacchi, C. Larizza, C. Combi, R. Bellazzi, Data mining with temporal abstractions: learning rules from time series, Data Min. Knowl. Discov. 15 (2) (2007) 217–247.

[35] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.

[36] M.Q. Stearns, C. Price, K.A. Spackman, A.Y. Wang(Eds.), SNOMED clinical terms: overview of the development process and project status, in: AMIA Symposium, American Medical Informatics Association, 2001.

[37] G.M. Weber, S.N. Murphy, A.J. McMurry, D. MacFadden, D.J. Nigrin, S. Churchill, I.S. Kohane, The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories, J. Am. Med. Inform. Assoc. 16 (5) (2009) 624–630.

[38] C.Y. Wu, C.K. Chang, D. Robson, R. Jackson, S.J. Chen, R.D. Hayes, R. Stewart, Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register, PLoS One 8 (9) (2013) e74262.

[39] R. Yancik, M.N. Wesley, L.A. Ries, R.J. Havlik, S. Long, B.K. Edwards, J.W. Yates, Comorbidity and age as predictors of risk for early mortality of male and female colon carcinoma patients, Cancer 82 (11) (1998) 2123–2134.

[40] J. Hippisley-Cox, C. Coupland, Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm, Br. J. Gen. Pract.: J. R. Coll. Gen. Pract. 62 (2012) e29–e37.