Artificial Intelligence in Medicine 69 (2016) 53–61

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer

Mark Hoogendoorn[a,b,*], Peter Szolovits[b], Leon M.G. Moons[c], Mattijs E. Numans[d]

[a] Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[b] Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
[c] Department of Gastroenterology and Hepatology, Utrecht University Medical Center, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
[d] Department of Public Health and Primary Care, Leiden University Medical Center, Hippocratespad 21, 2333 ZD Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 6 November 2015
Accepted 23 March 2016

Keywords:
Natural language processing
Predictive modeling
Uncoded consultation notes
Colorectal cancer

ABSTRACT

Objective: Machine learning techniques can be used to extract predictive models for diseases from electronic medical records (EMRs). However, the nature of EMRs makes it difficult to apply off-the-shelf machine learning techniques while still exploiting the rich content of the EMRs. In this paper, we explore the usage of a range of natural language processing (NLP) techniques to extract valuable predictors from uncoded consultation notes and study whether they can help to improve predictive performance.
Methods: We study a number of existing techniques for the extraction of predictors from the consultation notes, namely a bag of words based approach and topic modeling. In addition, we develop a dedicated technique to match the uncoded consultation notes with a medical ontology. We apply these techniques as an extension to an existing pipeline to extract predictors from EMRs. We evaluate them in the context of predictive modeling for colorectal cancer (CRC), a disease known to be difficult to diagnose before performing an endoscopy.
Results: Our results show that we are able to extract useful information from the consultation notes. The predictive performance of the ontology-based extraction method moves significantly beyond the benchmark of age and gender alone (area under the receiver operating characteristic curve (AUC) of 0.870 versus 0.831). We also observe more accurate predictive models by adding features derived from processing the consultation notes compared to solely using coded data (AUC of 0.896 versus 0.882) although the difference is not significant. The extracted features from the notes are shown be equally predictive (i.e. there is no significant difference in performance) compared to the coded data of the consultations.
Conclusion: It is possible to extract useful predictors from uncoded consultation notes that improve predictive performance. Techniques linking text to concepts in medical ontologies to derive these predictors are shown to perform best for predicting CRC in our EMR dataset.

1. Introduction

Electronic medical records (EMRs) are a valuable resource in the development of predictive models for diseases. The increasing level of integration of information from different caretakers into single EMR systems increases the possibilities even more. The step from an EMR to a predictive model is, however, far from trivial and requires dedicated processing techniques for a number of reasons: First of all, EMRs are typically ambiguous because different caretakers use different coding conventions. Furthermore, some information stored in the system might require background knowledge or context to be sufficiently usable in the development of predictive models (e.g. a raw lab value). Third, information stored in EMRs is of a highly temporal nature, whereas traditional predictive modeling techniques are unable to take advantage of this temporal dimension. Finally, not all EMR data is always coded; uncoded notes written by a physician are frequently seen as part of EMRs.

In previous research (cf. [1]) we have developed a preprocessing pipeline that includes components to handle the first three characteristics of EMRs, allowing for the application of off-the-shelf machine learning algorithms while benefiting from the

* Corresponding author at: Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.
E-mail addresses: m.hoogendoorn@vu.nl (M. Hoogendoorn), psz@mit.edu (P. Szolovits), l.m.g.moons@umcutrecht.nl (L.M.G. Moons), m.e.numans@lumc.nl (M.E. Numans).

http://dx.doi.org/10.1016/j.artmed.2016.03.003
0933-3657/

rich content of the EMRs. However, the pipeline does not yet include a natural language processing (NLP) component which is able to distill useful information from uncoded notes. Research has shown (see e.g. [2]) that such notes can be beneficial when it comes to the development of predictive models, even when coded data are present. In this paper, we study three different NLP approaches and investigate their added value: (1) a simple bag-of-words approach (seen as a benchmark), (2) a topic modeling approach using both latent dirichlet allocation (LDA, cf. [3]) and hierarchical dirichlet processes (HDP, cf. [4]) where topics of text descriptions are identified in an unsupervised way using Bayesian learning, and (3) a dedicated approach (introduced in this paper) which matches the text with a medical ontology (UMLS [5] and an alternative coding scheme called ICPC [6]). Although there are numerous studies aimed at extracting knowledge from medical text, hardly any has tried to compare a range of techniques. In addition, the notes we study are brief, and more keyword oriented and not full blown reports, making the application of known NLP tools for processing medical uncoded text (e.g. [7,8]) less appropriate.

We study the performance of the different NLP techniques in the context of a large anonymized primary care dataset we have access to, covering around 90,000 patients in the region of Utrecht, the Netherlands. Specifically, we focus on predictive modeling of colorectal cancer (CRC), a disease known for its nonspecific symptoms. The dataset consists of coded data (on lab measurements, diagnoses during consultations, medications, and referrals) and includes uncoded doctor's notes associated with each consultation/visit of a patient. We aim to answer the following questions:

1. Can we distill information from the consultation notes that has predictive value with respect to CRC, and if so, what NLP technique results in the highest benefit?
2. Do the consultation notes have added predictive value in addition to the coded data in the dataset?
3. Can we obtain enough information from the consultation notes by themselves to obtain at least equivalent predictive performance for CRC as from just the coded data associated with the consultation?

This paper is organized as follows. Section 2 gives an overview of related work. Thereafter, the dataset is described in more detail in Section 3 followed by an explanation of the different algorithms we use for processing the notes in Section 4. The experimental setup to evaluate the algorithms and answer the research questions is expressed in Section 5 whereas the results are presented in Section 6. Finally, Section 7 is a discussion.

2. Related work

A variety of studies have been performed to explore how useful information can be extracted from medical texts and how valuable this information can be in predictive modeling.

First of all, a number of tools have been developed that allow for the identification of medical terms from text, thereby coupling it to a medical ontology (typically UMLS or a subset thereof). Good examples of such tools are MetaMap [9], health information text extraction system (HITEx, cf. [8]), and cTAKES [7]. All these tools perform basic pre-processing operations first (e.g. tokenization, stemming) and then use an algorithm to perform the best matching with a medical ontology. Research in this area is mostly focused on getting a high accuracy, i.e. attributing the right terms from the medical ontology to the text. The tools aim at exploiting properties of rich, full sentences written in the English language.

Studies that explore the benefit of distilling information from the text (i.e. uncoded data) for the purpose of predictive modeling are more limited in number. In [2] a study is performed in the area of rheumatoid arthritis showing that typed physician notes can complement coded data and result in a higher predictive value. Here, the HITEx system was used to extract relevant UMLS terms from the notes. Ref. [10] studies the usage of topic modeling to help in the classification of pediatric patients, specifically in the prediction of infant colic and shows interesting insights that can be gained from the application of such techniques. In [11] topic modeling is applied on an ICU progress notes dataset to identify mortality risk for ICU patients. The program first assigns UMLS terms to the notes, followed by the application of topic modeling over those terms. The predictive performance is significantly improved compared to the performance without utilizing the notes. Ref. [12] describes a topic modeling approach for mortality prediction based on free-text hospital notes. They have developed models for different time windows, namely in hospital, 30 day post-discharge, and 1 year post discharge. Their results show an improved predictive value in case the text notes are processed using topic modeling in all different time window settings. Finally, Luo et al. [13] explores the usage of more information from the structure of sentences by exploring them in the form of graphs where the nodes in the graph are medical terms found in the sentence and the edges involve the role of the word in the sentence. Subgraphs are identified that occur frequently. The system is shown to outperform all benchmarks. As can be seen, most predictive modeling approaches focus on one technique and try to show that the predictive power is improved; however none study a wide range of techniques and their individual contributions or benefits.

The studies described above focus on one specific technique to distill information from text. In very few works, a comparison of different approaches to extract knowledge from text is found. Tremblay et al. [14] is however an exception: the authors try to explore whether both supervised and unsupervised learning methods can be used to enhance coded data (that might be incomplete). The study focuses on fall injuries. Overall, they show that the two different types of approaches could complement the coded data.

There are also various studies that aim at predicting the specific disease that is the subject of our study: colorectal cancer. Two models in particular are worth mentioning: the Bristol Birmingham equation [15] and the model by Hippisley-Cox [16]. Both have been generated using primary care data. We have used the Bristol Birmingham equation as a benchmark before and have shown that we are able to move statistically significantly beyond that model (cf. [1]). In this paper, we will merely focus on the benefit of using the uncoded notes compared to the performance we have already obtained.

3. Dataset description and preparation

We analyzed an anonymized primary care dataset originating from a network of general practitioners (GPs) centered around the Utrecht University Medical Center, the Netherlands. It contains data of a total of just over 90,000 patients[1] for the period between July 1, 2006 and December 31, 2011. The number of positive CRC cases in the dataset is 588. The dataset covers the following information for each patient, all stored by the date at which the activity took place:

[1] Note that previously [1] we have reported on datasets covering more patients. The dataset we are using for this research is more limited in terms of number of patients, as we only have access to the consultation notes for a subset of the dataset we reported on earlier. It concerns a part of the previously reported dataset covering the practices working with the Promedico ASP system. Experiments using merely coded data do not show considerable differences in performance on this subset compared to the dataset covering more patients.

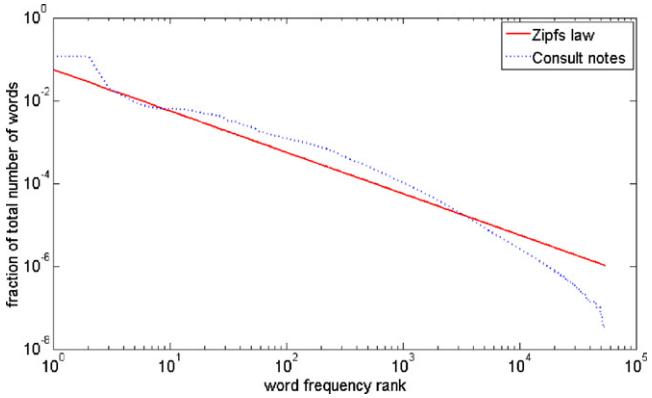

Fig. 1. Frequency of words, starting with the most common word compared to Zipf's law. Note the log scale on both axes.

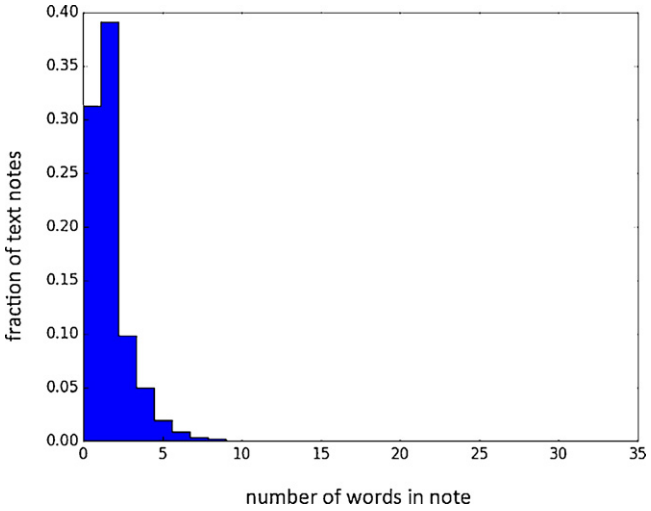

Fig. 2. Length distribution of consultation notes.

- *Consults:* code of symptoms and/or diagnoses during the consultation/patient visit, assigned according to the ICPC coding standard [6], using the Dutch version.
- *Medication:* medication prescribed, including the dosage. The coding scheme applied for medication is the ATC scheme [17].
- *Lab results:* any form of lab measurement that has been performed by the GP, or that was received from an external lab. The coding scheme is specific for the GP information system from which the dataset was exported.
- *Referrals:* referrals to secondary care, again coded in an information system specific way.
- *Consultation notes:* uncoded notes entered by the GP associated with consultations, written in Dutch.

In order to prepare the dataset for learning, a total period of six months of relevant data has been selected per patient. Patients that were not enrolled for a sufficiently long period have been removed from the dataset. For CRC patients, the six month period prior to diagnosis is used, following previous work (cf. [15,1,18]). We exclude any data on the last two days before the actual diagnosis. For other patients, a random period of six months has been selected (cf. [15]). Only patients aged 30 and up have been included, as the likelihood of developing CRC below 30 is extremely low. The dataset (after selection) is sizable: a total of 502,000 consultations evaluated with a coding are available, 344,000 medication prescriptions, 316,000 grounded lab result records (see Section 4), 2800 referrals, and 1.25 million consultation notes. When considering these numbers, most of them are in line with our expectations except for the limited number of referrals. Apparently the dataset did not include this information in full. The number of patients and CRC cases specified above (i.e. 90,000 and 588) are those in the dataset after selection.

Because the main emphasis of this study is to explore the value of the consultation notes, we will explore these data in a bit more detail. This analysis is based on the full set of notes (i.e. not the selections of six months nor the selection based on age), covering over 14.6 million notes. Fig. 1 shows the frequencies observed for words in the notes. It can be seen that the distribution is highly skewed and does not completely follow Zipf's law [19]. Notes in this dataset are rather brief (shown in Fig. 2), most likely due to the type of interaction GP's have with their patients: in general these are short visits and the various codings provide quite a natural way of storing the information associated with the visit. The notes are usually a (limited) expansion of the textual description of the code assigned to the visit.

4. Methodology

In this section, the pipeline developed to pre-process the EMR data in a suitable way will be discussed. First, the initial version of the pipeline (as reported in [18]) will briefly be described, followed by the natural language processing extension investigated in this paper.

4.1. Initial processing pipeline

Processing the data described in Section 3 is far from trivial. Although most of the data is structured in a nicely coded format, simply applying an off-the-shelf machine learning algorithm would not provide satisfactory results. In order to fully profit from the wealth of information residing in the data, a pipeline has been developed in earlier research [1] and is shown in Fig. 3. It goes beyond the scope of this paper to discuss the whole pipeline in detail, but it consists essentially of three main steps, which are briefly described below.

1. *Lab result contextualization:* lab results require interpretation of their raw values to be meaningful. This step determines whether a value is within the normal range, low or high (referred to as *absolute grounding*), and interprets it compared to previously observed values, thereby identifying a trend: stable, increasing or decreasing (i.e. *relative grounding*). To limit the number of trends to be considered, only those trends are selected that are sufficiently supported in the data, as described in the *Temporal Patterns Identification* step, below. Note that absolute grounding can only be derived if the normal ranges are available for the specific type of measurement.
2. *Semantic enrichment:* there are often biases when it comes to the registration of data related to a patient in the information system, mostly caused by lack of time. The idea of the semantic enrichment step is to use the available data and enrich it based on domain ontologies (e.g. SNOMED [20]). For instance, if a certain type of medication is prescribed, the indication accompanying that type of medicine can be added to the dataset. Due to the disappointing performance of this step (see [1]) we have however decided to exclude it in the experiments reported in this paper.
3. *Temporal pattern identification:* temporal developments with respect to the health state of a patient are crucial, for instance to identify whether symptoms worsen over time. Traditional machine learning techniques are not able to grasp these patterns.

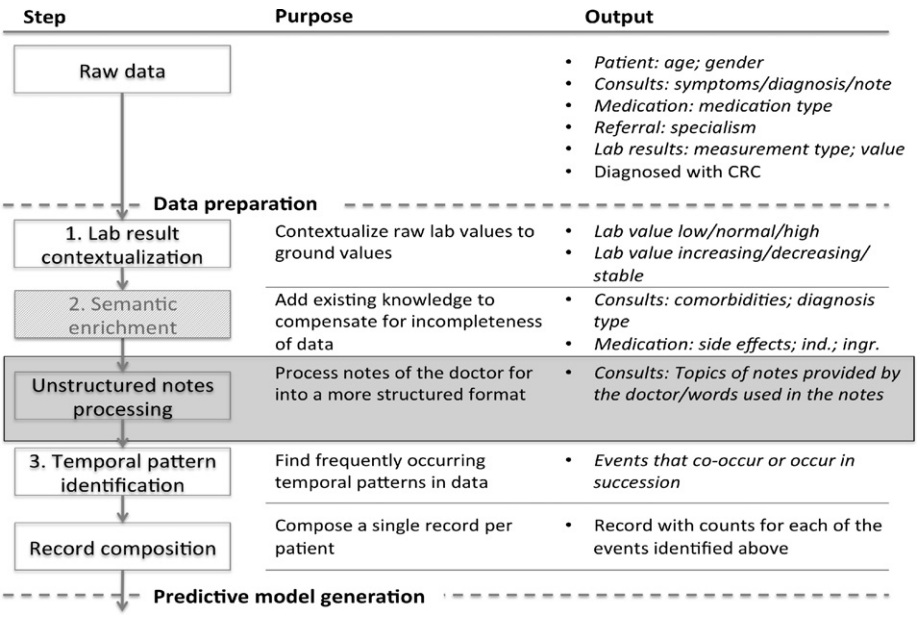

Fig. 3. The pipeline to process the GP dataset. Grey indicates the main contribution of this paper.

Therefore, in this step frequently seen co-occurrence and succession of events patterns are derived from the data. To generate these patterns in an effective way, a dedicated algorithm inspired by [21] has been developed, where we only use a subset of their approach: merely consecutive events that occur in sequence. Only those temporal patterns are generated that fulfill a minimum support threshold in the data (i.e. they occur for a fraction of either CRC or non-CRC cases that is above this minimum support threshold) and show a statistically significant difference in occurrence between the two groups. Such a selection is needed to limit the huge number of possible temporal patterns. See [1] for more details.

4.2. Natural language processing extension

The main contribution of this paper comes from the addition of an extra step in the pipeline, namely the processing of the uncoded notes, in this specific case associated with consultations. Previous work, as identified in Section 2, has shown that these notes can be of added value and bring substantial predictive power. Given the nature of the notes in our dataset, which are short, keyword-like pieces of text, we have decided to focus on approaches that do not try to exploit the structure of the sentences in the text. Instead, we focus on topic modeling and matching with medical ontologies.

Fig. 4 shows the NLP pipeline we have developed. The left side shows a common approach to handling text by first performing tokenization, followed by stemming of the words identified in the text. Each element in the dataset is one note written by a GP associated with a single consultation of a patient. Based on these notes, we identify attributes that can be used in our eventual learning set. We distinguish six different approaches, each treated in more detail below. Note that the first three approaches are off-the-shelf approaches while the last three approaches have been developed specifically in this paper. We have selected these techniques because they cover the entire range of NLP techniques: a benchmark (bag of words), unsupervised methods to extract information from text (topic modeling), and specifically designed approaches for the case at hand (the remaining approaches). Our choices for specific techniques within these categories are based on observations from the literature.

1. *Bag of words:* the simplest way to generate attributes is to identify all unique words that are present in the total set of all notes. Stop words are removed because they are highly unlikely to have predictive value and might distort the machine learning algorithm. For each individual note, the values for the attributes are the number of occurrences of the word in the consultation note.
2. *Topic modeling with balanced notes:* the second approach is a topic-modeling approach, applied to the stemmed notes as they are. Two different approaches are used: the parameterized LDA approach (cf. [3]), requiring a desired number of topics to be set, and the non-parameterized HDP approach (cf. [4]) which finds the best number of topics itself. Both use a form of Bayesian learning whereby words that tend to co-occur form a topic. The approaches are unsupervised and try to extract more high-level concepts from all of the notes (i.e. the topics). They do not take into account whether the patients reside in the CRC or non-CRC group as they do not aim to make predictions directly, they essentially derive potentially useful attributes that can be used in our learning algorithm. The topics are specified by the words that are covered by the topic and a weight given to each of the words. One aspect that deserves attention is the highly unbalanced nature of the dataset: only 588 out of a total of over 90,000 patient are diagnosed with CRC. If we were to run a topic modeling approach on the whole dataset it would most likely not find interesting predictors as the data of the 588 (for our case highly interesting) patients will not stand out compared to the huge number of patients that do not develop CRC. Therefore, we oversample the notes from the CRC patients to get a more even ratio between them and the non-CRC group; the precise value selected is specified in Section 5. Then, we apply one of the two aforementioned topic modeling approaches. Once the topics have been identified, notes can then be scored against these topic descriptions, and the topic with the highest score is assigned to the note. An alternative would have been to include the score on (a subset of) the topics for each note, but we feel that this choice was in line with the limited amount of text seen in the notes.
3. *Topic modeling for two classes separately:* although the approach to oversample data from the CRC cases will most likely result in more relevant predictors being found, it does not guarantee that the topics will cover the unique features of the notes associated with the CRC patients. Therefore, a second approach is taken

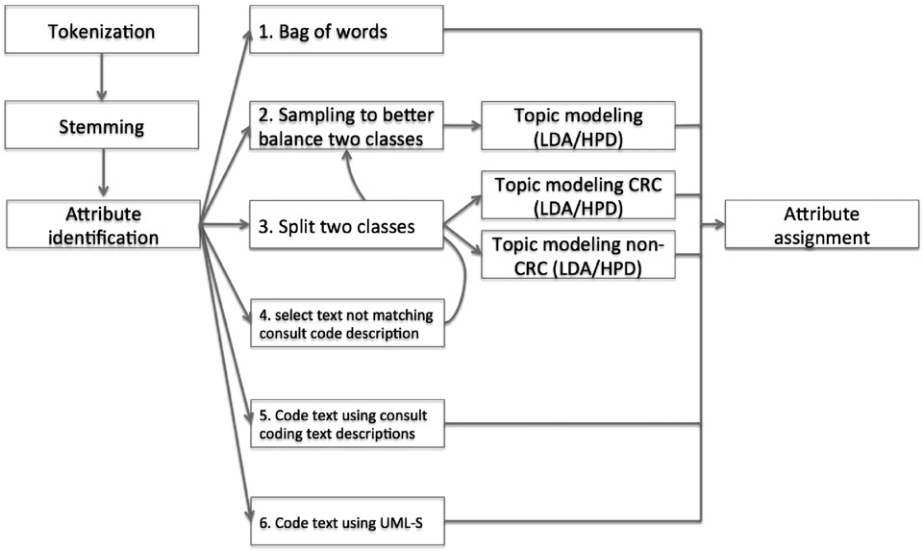

Fig. 4. The text processing pipeline on the left, where the different variants of the attribute identification stage are shown in the middle, using which new attributes and their accompanying values are assigned (right).

that creates separate topic models for CRC and non-CRC patients. Each note is scored against the two models and the topic scoring highest (independent of its source) is assigned to the note.

4. *Topic modeling for text with potentially added value:* one of the main goals of this paper is to explore whether the text of the consultation notes brings *additional* predictive value. When we examine the notes, we see that often they contain a phrase that is highly similar to the description of the ICPC code that has been assigned to the consultation. While this might not necessarily be a problem, we are interested in seeing whether the text contains additional terms for concepts that are not covered by the coding. Therefore, we have also developed a variant in which we filter the terms in the text based on the textual description of the code assigned to the particular consultation. We calculate the Levenshtein distance (cf. [22]) between the stemmed words that are part of the note and the stemmed keywords that are part of the description of the code associated with the consultation. For each word where there is at least one keyword having a sufficiently low Levenshtein distance, the word is removed from the specific note and thus not considered when applying the topic modeling approach. After this filtering, the same steps are performed as described under the topic modeling approach with balanced data (i.e. approach 2).
5. *Coding the notes using textual descriptions of consultation codes:* in order to study whether the consultation notes (which mainly concern consultation descriptions) can be coded using a consultation coding scheme, we apply a reasonably simple scheme very much in line with finding the text with added value: we look at the stemmed keyword descriptions accompanying the ICPC codes and the stemmed words as part of the note, and measure the Jaccard similarity (cf. [23]) between the two. To ease the computational complexity (1.25 million consultation notes, 685 codes) we apply a two phased approach. In a first (rough) matching we make a pre-selection of ICPC codes by selecting those codes that have at least one keyword that is equal or a superstring of at least one stemmed word in the consultation note. For the selected ICPC codes we measure the Jaccard similarity in the second phase. Here, words are matched on a more fine-grained level (not just focusing on substrings) by means of the Levenshtein distance. Words with a distance of one or less are considered a match and thus part of the intersection. We then assign the best scoring ICPC code (i.e. with the best score on the Jaccard similarity) to the consultation note. In case of multiple competing codes with the same score, the code with the lowest alphabetical ICPC code is selected. If no code could be found with at least one match, the general consultation code is assigned to the note.
6. *Coding the notes using UMLS:* although the ICPC code matching could be a viable option, it clearly does not have the richness of an ontology such as UMLS. Therefore, we also try a more sophisticated approach: We match words in our text notes with the descriptions of concepts that are part of SNOMED-CT (which has a Dutch text description of medical terms), then link these concepts to a UMLS Concept Unique Identifier (CUI). For each word in our note, we search for all CUI's for which the description contains our words or their superstrings. For each word, we obtain this list of CUI's and look for commonalities between the lists of the different words in a single note. The CUI that occurs most frequently among the joint set of words is selected. In case there are multiple competing CUI's, previously assigned CUI's are preferred. If multiple competing CUI's are still present, the one with the alphabetically lowest CUI is selected. If not a single matching CUI is found, the value *unknown* is assigned. Note that we did not use one of the known parsers described in Section 2 because the text is of a completely different nature (very keyword-oriented and of very limited length) and, in addition, is in Dutch.

5. Experimental setup

This section describes the experimental setup that has been used to answer the research questions we have posed related to the consultation notes data. First, the machine learning setup will be explained as well as the various parameter settings. Then, we will describe the setup for the experiments that allow us to answer the research questions.

5.1. Machine learning setup and parameter settings

For this paper, we do not aim to study which machine learning technique works best, but focus on the benefit of using the consultation notes. As a result, we will only use one algorithm, and have chosen the one that showed the most promising performance in combination with the pipeline in previous work [1]: logistic regression. Mean squared error (MSE) is used as the evaluation

criterion. Before applying logistic regression we apply feature selection using the Pearson correlation coefficient, thereby selecting the 50 features that were shown to have the strongest correlation (be it negative or positive) with the target. We did not take correlations between the attributes (or features) into account. Previous work has shown that this number boosts performance compared to a higher number of attributes and still leaves room for finding new predictors compared to known ones. In order to evaluate the different setups we use 5-fold cross validation and measure the area under the receiver operating characteristic (ROC) curve averaged over the 5 folds. We abbreviate this to the AUC. In addition, we measure the precision, recall, and F1 measure. When studying the final predictors, we generate a model over the entire set of patients (unlike the cross-validation, where we use five configurations of a training and test set) using the setup that generated the model that showed the best performance in the cross validation. For temporal pattern generation, we set the minimum support of patterns (to be met for either CRC or non-CRC patients or both) to 0.1 for the entire set of patients, and 0.15 for HDP topic modeling and text matching using ICPC and UMLS matching to sufficiently limit the number of patterns generated and not to run into computational problems. Note that given these constraints, we still find a reasonable number of patterns, mainly among the CRC patients; they tend to be a more homogeneous group than the controls from the entire non-CRC population, and thus exhibit more common patterns.

For topic modeling, we have set the number of topics for LDA to 50. We have tried more topics, but it did not improve performance. Furthermore, for similarity scores between words to identify unique text compared to the coded data, words with a Levenshtein distance of 2 or less were considered too similar. For the text modeling variants that require an oversampling ratio to be set, we set the desired CRC to non-CRC ratio to 0.25, thereby avoiding dominance by the limited number of CRC patients while still making their data substantial enough to be a prominent part of the topic modeling. Finally, for the UMLS matching, a minimum word length of 5 has been imposed; shorter words proved to distort the learning process as too many CUIs match, resulting in large inaccuracies. Experiments with a minimum word length of 3 provided substantially worse results.

5.2. Experiments

In order to answer our research questions, we performed a number of experiments using different parts of the dataset.

For our first set of questions, we aim to study whether the text data brings additional predictive value. For this purpose, we always use the age and gender (AG) of the patients, and extend this with either only the text data (T) for the patient or we extend it with all coded data (consultation codes (C), medication (M), referrals (R), and lab results(L)) combined with the text data. For these different subsets, we apply the pipeline as explained in Section 4. To understand the benefit of using temporal patterns, we also look at the difference when using plain counts, and additionally using the temporal pattern identification step. For a comparison of the AUC compared to currently existing predictive models, see [18].

Next we want to judge whether the consultation notes could replace part of the coded data, namely the consultation coding (C). Of course we are mainly interested in the added value of the notes on top of the coded data, but it is still worthwhile to explore how much the two differ in terms of predictive content to gain good insights. To answer this question, we omit the coded consultation data from the learning sets identified in the previous setup and apply the same pipeline again. Here, we only take the most promising text processing techniques identified in the previous set of experiments.

6. Results

This section describes the results we have obtained using the aforementioned experimental setup. First, we show the benefits of using the consultation notes and present an in-depth analysis. Then we examine the ability of features derived from the consultation notes to substitute for coded data.

6.1. Benefit of using consultation notes

The results using the different methods explored to extract predictors from the text are shown in Table 1. Note that these results are the average AUCs obtained in the 5-fold cross-validation approach. The table also includes the 95% confidence intervals, which enables a statistical comparison between pairs of the reported AUCs.

From the table, it can be seen that all text processing techniques are able to extract some information from the text that is usable as a predictor, except for the bag of words approach. Below, we also express the precision, recall, and F1-scores obtained for the most promising experiments, as the AUCs can sometimes be misleading for highly unbalanced data.

When solely using the text data in combination with age and gender (that are by themselves already quite predictive), we move up from an AUC of 0.831 (precision: 0.65, recall: 0.023, and the F1-score: 0.045) to at most 0.865 (precision: 0.029, recall: 0.65, F1-score: 0.055) for the UMLS based approach, which clearly has the highest score. Do note that the difference is not statistically significant: due to the limited number of CRC cases, the 95% confidence interval is relatively wide, resulting in few statistically significant differences in performance. Different variants of the topic modeling approaches do not exhibit very distinct performance. The matching of the text using the description of the ICPC coding scheme performs a bit better than topic modeling, but substantially worse than the UMLS based approach.

When we focus on the increase of performance when the derived text attributes are used in combination with all coded data, we see that the topic modeling approaches do not identify new valuable predictors in addition to the coded data. The UMLS based approach does however increase AUC performance substantially: 0.896 (precision: 0.035, recall: 0.68, F1-score: 0.067) compared to 0.882 (precision: 0.031, recall: 0.69, F1-score: 0.060) without the text data. Identifying the unique text in addition to the coded data does not show the expected benefits, nor does the coding using the ICPC text description scheme. Once temporal patterns are identified and used, more accurate models are created. The UMLS-based approach still performs best (0.900 with precision: of 0.039, recall: 0.70, F1-score: 0.074), although the difference from the other approaches is slightly smaller (0.01). This shows that the temporal patterns are in general more robust in finding good predictors and less dependent on all different categories in the data.

To summarize: yes, the consultation notes contain useful information that can be extracted best using the UMLS based approach, and we see signs that this can move the predictive performance beyond the coded data. We base the first 'yes' on the fact that the confidence interval of the setup with the UMLS approach combined with age and gender does not overlap with the confidence interval of age and gender alone. The second observation cannot be answered with a definitive 'yes' as the performance does not move significantly beyond the performance of the coded data.

6.2. In-depth analysis

Of course, merely looking at predictive performance is not the only aspect of interest; we have also analyzed what information is being extracted from the consultation notes. Here, we focus on

Table 1
Added value of using the consultation notes. AUCs and 95% CIs for all combinations. For the data, AG = age/gender; C = consultation code; M = medication; R = referrals; L = lab results, and T = text/consultation notes. The numbering convention used for the approaches follows Section 4, 1 = bag of words; 2 = topic modeling with oversampling; 3 = separate topic modeling for two classes; 4 = topic modeling for text beyond consultation code; 5 = coding using ICPC descriptions, and 6 = coding using UMLS.

Text processing technique	Data	Regular counts	Temporal patterns plus regular counts
–	AG	0.831 (0.814–0.848)	0.831 (0.814–0.848)
1	AG/T	0.825 (0.808–0.842)	0.826 (0.809–0.843)
2-LDA	AG/T	0.848 (0.831–0.865)	0.854 (0.838–0.870)
2-HDP	AG/T	0.847 (0.830–0.864)	0.847 (0.830–0.864)
3-LDA	AG/T	0.846 (0.829–0.863)	0.852 (0.836–0.868)
4-LDA	AG/T	Not relevant to study in isolation	
5	AG/T	0.853 (0.837–0.869)	0.855 (0.839–0.871)
6	AG/T	0.865 (0.849–0.881)	0.870 (0.854–0.886)
–	AG/C/M/R/L	0.882 (0.867–0.897)	0.890 (0.875–0.905)
1	AG/C/M/R/L/T	0.882 (0.867–0.897)	0.890 (0.875–0.905)
2-LDA	AG/C/M/R/L/T	0.885 (0.870–0.900)	0.890 (0.875–0.905)
2-HDP	AG/C/M/R/L/T	0.886 (0.871–0.901)	0.890 (0.875–0.905)
3-LDA	AG/C/M/R/L/T	0.884 (0.869–0.899)	0.890 (0.875–0.905)
4-LDA	AG/C/M/R/L/T	0.884 (0.869–0.899)	0.890 (0.875–0.905)
5	AG/C/M/ R/L/T	0.886 (0.871–0.901)	0.890 (0.875–0.905)
6	AG/C/M/R/L/T	0.896 (0.882–0.910)	0.900 (0.886–0.914)

Table 2
Overview of top predictive topics using the 2-LDA. The left column describes the stemmed words (in Dutch), followed by an expert interpretation of the terms. The rightmost columns show the weight the topics obtained in the final logistic regression equations that resulted from that specific setting (Table 1 fourth row, third column for the *Weight AG/T* column and the twelfth row, third column for the *Weight AG/C/M/R/L/T* column). Note that merely one topic is present in the final equation with the coded data.

Key topic terms, stemmed (top 2)	Expert description	Weight AG/T	Weight AG/C/M/R/L/T
Coloncarcinom (0.22), malais (0.12)	Suspicion of CRC	0.067	0.337
Algemen (0.50), journal (0.49), brak (0.002)	General journal text	0.058	–
Anemie (0.16), chronisch (0.12)	Anemia	0.051	–
Met (0.18), rectumcarcinom (0.07)	Suspicion of CRC/cancer	0.048	–
Malign (0.10), duizel (0.10)	Suspicion of cancer	0.043	–

the variant of the pipeline without the temporal patterns, as this shows the biggest improvement when studying the notes and also provides more insight compared to analyzing the more complex patterns used in the temporal approach.

Topic modeling: first, we consider the best performing topic modeling approach (although the differences are very small): LDA in combination with oversampling of the CRC cases (i.e. the 2-LDA rows in Table 1). The most important predictors found (for the non-temporal case, the 3rd column in Table 1) are shown in Table 2.

From the table, it can be seen that topics are identified that relate to a suspicion of CRC or cancer. Note that this concerns a suspicion, it is not coded as such as the note would otherwise not be part of the selected notes since we set the end point to two days before the CRC coding. In addition, general consultations seem predictive, as is anemia, which is in fact a known predictor. In combination with coded data, only the topic related to the suspicion of CRC is included among the selected top 50 predictors, showing that their predictive performance on top of the coded data is limited.

ICPC matching: second, we have studied the performance of the scheme to match the consultation notes to the coding scheme used for consultations (ICPC). Specifically, we study how much the resulting ICPC codes coincide with the code assigned to the consultation. Table 3 shows the results.

The overall accuracy of the precise matching is quite poor: only 37.3% of the notes is assigned to the proper code. Of course, the notes do not necessarily have to concern a remark that coincides with the code. Some categories score a bit higher, whereas mainly the general (A) category scores very low. This is explainable as this is a category under which general complaints of great diversity fall. If we look at the matching with the code group, we see a more reasonable (though not great) performance.

UMLS matching: finally, we study the most promising model, whereby we match the text with UMLS and assign the most likely CUI to the consultation note. Table 4 shows the ten CUIs that are used in the eventual non-temporal model with all coded data.

Table 3
Performance analysis of simple coding scheme for coding the consultation notes (i.e. alternative 5). The overall, best and worst three categories are shown as well as those relevant for CRC that are not among the best/worst performers. The percentages are computed over cases attributed to the specified code group by the doctor. A case was counted correct if precisely the same code was assigned (column 2) or in case the code group was correctly identified (column 3).

Code group	Accuracy – precise match	Accuracy – code group match
All	37.3%	49.0%
Best three		
H (ear)	63.1%	68.1%
T (endocrine/metabolic and nutritional)	59.2%	60.1%
R (respiratory)	53.9%	59.3%
Worst three		
X (female genital)	20.6%	31.5%
Z (social problems)	16.7%	37.2%
A (general and unspecified)	8.6%	29.8%
Additional CRC relevant		
D (digestive)	45.0%	58.4%

From Table 4, it can be seen that the suspicion of CRC and cancer is also reflected in the CUIs identified, although these are not identified frequently among the consultation notes of patients. In addition, some relevant other predictors are identified that are known to be predictors for CRC: increased bowel frequency, anemia, and sanitary drainage.

In order to understand the difference in performance a bit better, we have explored for what patients the models with and without the use of the consultation notes make different predictions (assuming the point on the ROC curve) where the false positive rate is around 40% where the curve flattens). When we fix the threshold of the logistic regression equation to accomplish that rate for

Table 4
Description of most predictive UMLS terms in the model using technique 6 when using all coded data without temporal patterns. Note that for the logistic regression, independence of attributes has been assumed when constructing the logistic regression model (i.e. no interaction terms have been considered).

CUI	Textual description	Weight AG/C/M/R/L/T	Number of notes (patients)
C0239978	Increased bowel frequency	0.76	39(9)
C0279639	Mucinous adenocarcinoma of the colon	0.67	81(3)
C0007113	Rectal cancer NOS	0.58	785(105)
C0037073	Neoplasm, sigmoid	0.35	119(7)
C0085576	Anemia microcytic	0.34	1563(214)
C0013106	Drainage, sanitary	0.312	533(69)
C0006430	Burning mouth syndromes	0.276	85(16)
C0751498	Cancer of sigmoid	0.228	39(5)
C0684337	Neuroectodermal tumor, peripheral	0.016	40(2)
C0861772	Rectal cancer Duke's D	−0.047	24(3)

Table 5
Overview of CRC patients that have been classified differently between the predictive model utilizing the consultation notes versus the model only using the coded data. Only the patients which score on one of the distinct attributes of the two models have been shown and only the score on those distinct attributes are shown.

Patient	Classified correctly with/without text from consultation notes	Age	Patient attributes (italics indicates unique predictor for non-text model, bold for text)
73109	With text	38	*C: flu vaccination = 1, L: normal creatinine = 1, L: normal hemoglobine = 2, L: normal mean corpuscular hemoglobin = 2, L: increasing erythrocytes sedimentation rate = 1*
77605	Without text	45	*C: flu vaccination = 1*
78583	With text	40	*L: normal creatinine = 1*
205419	With text	53	**T: suspicion rectal cancer = 4**
208088	Without text	69	*C: flu vaccination = 1, L: normal creatinine = 2, L: normal hemoglobine = 3, L: normal mean corpuscular hemoglobin = 3*
210563	With text	48	*L: normal hemoglobine = 1, L: normal mean corpuscular hemoglobin=1*, **T: suspicion rectal cancer = 9**
231309	Without text	63	*C: flu vaccination = 3, L: normal creatinine = 2, L: normal hemoglobine = 2, L: normal mean corpuscular hemoglobin = 2*

both models, 865 patients are categorized differently by the two algorithms. Among these patients are 10 CRC patients, where 7 are classified correctly with the text attributes and 3 without the text attributes. Out of the latter group of patients, Table 5 shows those seven patients for which at least one unique predictor from the two models had a non-zero value.

Table 5 shows that the suspicion for CRC expressed in the consultation notes is quite beneficial in the predictive modeling of the seven cases: for two out of seven the physician has entered such a suspicion a number of times. Overall, there are quite a substantial number of relatively young patients among the group, most likely causing them to be more difficult to classify. The flu vaccination (not considered in the model with text) seems to contribute to a better prediction for the non-text model in some cases. This can be explained by the fact that it acts as a proxy for comorbidity and general weakness as these are among the criteria to provide such a vaccination. Finally, it can be seen that the non-text model uses a lot more lab measurements, which do not necessarily contribute to a more accurate prediction.

6.3. Substitutability of coded data with consultation note information

The final subject of our study is to explore whether the consultation notes can replace the coded data related to consultations. This allows us to understand how the two relate to each other in terms of predictive value and content. To answer this question, we compare the best topic modeling approach (LDA with oversampling of CRC patients), the best approach that matches the note with codes (UMLS) combined with either only age and gender or with all coded data excluding the consultation codes, and a setting where the consultation codes are used while the consultation notes are not. Table 6 shows the results.

The results show that the topic modeling approach can certainly not substitute for the coded data. However, when using the UMLS matching scheme, we see that the approach, when excluding the consultation codes, even performs better than the set where the consultation codes are used while the text attributes are not (0.887 versus 0.882, though the confidence intervals overlap). Given our

Table 6
Ability of the consultation notes to replace the coded consultation data. AUCs and 95% CIs for all combinations. For the data, AG = age/gender; C = consultation code; M = medication; R = referrals; L = lab results, and T = text/consultation notes.

Text processing technique	Data used	Regular counts	Temporal patterns plus regular counts
–	AG	0.831 (0.814–0.848)	0.831 (0.814–0.848)
–	AG/C	0.879 (0.864–0.894)	0.881 (0.866–0.896)
2-LDA	AG/T	0.848 (0.831–0.865)	0.854 (0.838–0.870)
6	AG/T	0.865 (0.849–0.881)	0.870 (0.854–0.886)
–	AG/M/R/L	0.861 (0.845–0.877)	0.876 (0.861–0,891)
–	AG/C/M/R/L	0.882 (0.867–0.897)	0.890 (0.875–0.905)
2-LDA	AG/M/R/L/T	0.867 (0.851–0.883)	0.877 (0.862–0.892)
6	AG/M/R/L/T	0.887 (0.872–0.902)	0.893 (0.879–0.907)

earlier observations, we can conclude that the consultation notes contain predictors that have similar predictive power as the coded consultation data. The fact that predictive performance is even higher when the two are combined shows that they are different. This conclusion does not hold when the other coded data (i.e. medication, referrals and lab results) are excluded. Apparently the information extracted from the notes truly complements and does not substitute for those data. If temporal patterns are used, the performance difference is a bit smaller (0.893 versus 0.890). In general exploiting the temporal dimension shows a bit more robust performance when certain parts of the data are left out.

To summarize this aspect: yes, the consultation notes can replace the coded consultation data as the performance does not differ in a statistically significant way.

7. Discussion

In this paper, we have studied the use of consultation notes in predictive modeling of colorectal cancer in the context of a generic pipeline to pre-process EMRs in a suitable way and get accurate predictive modeling performance. To this end, we have applied a number of techniques that allow us to extract useful attributes from consultation notes, ranging from unsupervised learning techniques (i.e. different variants of topic modeling) to more knowledge driven approaches whereby we have introduced novel algorithms to code the information contained in the consultation notes using the coding convention used in the dataset (ICPC) or by matching with UMLS terms.

The results show that all techniques are able to distill interesting attributes that have predictive value, but when it comes to a true benefit compared to coded data, only the more knowledge-driven approaches make a real difference. The UMLS-based approach was even shown to make a substantial improvement in predictive performance on top of the coded data when simple counts of occurrences of attributes are used. In an endeavor to study whether the text notes contain as much predictive information as the coded equivalent, we did see that the predictive performance is indeed the very similar (again with the UMLS approach). An important remark in this context is that the combination of age and gender is already highly predictive for CRC, making the benefits of using additional predictors relatively modest. One important note that we need to make involves the language used in the notes: Dutch. Due to the more limited number of training corpora, stemming for the Dutch language is less accurate than e.g. English. Furthermore, only a small subset of UMLS (SNOMED-CT) is available in the Dutch language. Hence, the accuracy of the extraction of useful attributes may have been hampered due to these limitations.

While we are satisfied with the results obtained so far, there are ample avenues for future work. First, we want to study other diseases and see how well the results obtained for this setting generalize across diseases. Secondly, we want to study datasets originating from other GP information systems to explore whether our results generalize across the different systems.

Acknowledgements

We thank the GPs from the Julius General Practitioners Network (JHN) in the Netherlands, who were willing to share their anonymized routine general practice care EMR datasets for the purpose of this study. Prof. Szolovits' work is supported by NIH grants R01-EB017205 and 154HG007963. In addition, we would like to thank Reinier Kop for providing support with the integration of the developed text modules in the software accompanying the pre-processing pipeline and Tristan Naumann for the fruitful discussions.

References

[1] Kop R, Hoogendoorn M, ten Teije A, Buechner FL, Slottje P, Moons LMG, et al. Improved predictive modeling of colorectal cancer using a dedicated machine learning pipeline; 2015 [submitted for publication, available upon request].
[2] Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res 2010;62(8):1120–7.
[3] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
[4] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. J Am Stat Assoc 2006;101(476).
[5] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Suppl. 1):D267–70.
[6] Bentsen BG. International classification of primary care. Scand J Prim Health Care 1986;4(1):43–50.
[7] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.
[8] Zeng T, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inf Decis Mak 2006;6(1):30.
[9] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA symposium. American Medical Informatics Association; 2001. p. 17.
[10] Salleb-Aouissi A, Radeva A, Passonneau R, Tomar A, Waltz D, et al. Diving into a large corpus of pediatric notes. In: Proc. ICML workshop on learning from unstructured clinical text. 2011.
[11] Lehman L, Saeed M, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In: AMIA annual symposium proceedings, vol. 2012. American Medical Informatics Association; 2012. p. 505.
[12] Ghassemi M, Naumann T, Joshi R, Rumshisky A. Topic models for mortality modeling in intensive care units. In: ICML machine learning for clinical data analysis workshop. 2012. p. 1–4.
[13] Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc 2014;21(5):824–32.
[14] Tremblay MC, Berndt DJ, Luther SL, Foulis PR, French DD. Identifying fall-related injuries: text mining the electronic medical record. Inf Technol Manage 2009;10(4):253–65.
[15] Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut 2011;60(9):1242–8.
[16] Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract 2012;62(594):e29–37.
[17] World Health Organization, et al. Guidelines for ATC classification and DDD assignment. World Health Organization; 1996.
[18] Kop R, Hoogendoorn M, Moons L, Numans ME, ten Teije A. On the advantage of using dedicated data mining techniques to predict colorectal cancer. In: Holmes JH, Bellazzi R, Sacchi L, Peek N, editors. Proceedings of the 15th conference on artificial intelligence in medicine (AIME 2015). Springer; 2015.
[19] Zipf GK. Human behavior and the principle of least effort. Addison-Wesley Press; 1949.
[20] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the AMIA symposium. American Medical Informatics Association; 2001. p. 662.
[21] Batal I, Valizadegan H, Cooper GF, Hauskrecht M. A temporal pattern mining approach for classifying electronic health record data. ACM Trans Intell Syst Technol 2013;4(4):63.
[22] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Doklady 1966;10(8):707–10.
[23] Jaccard P. The distribution of the flora in the alpine zone. New Phytol 1912;11:37–50.