Text Mining to Support Knowledge Discovery from Electronic Health Records

Muhammad Zubair Afzal

The research presented in this thesis was financially supported by Miriam Sturkenboom's VICI award Innovation Incentive Scheme no. 91896632 from the Netherlands Organization for Health Research and Development ZonMw.

Financial support for publication of this thesis was generously provided by:

Erasmus University Rotterdam

Interdisciplinary Processing of Clinical Information (IPCI) group of the Department of Medical Informatics, Erasmus MC

ISBN: 978-94-6332-369-7

Layout: Ineke Jansen & Zubair Afzal

Printed by: GVO drukkers & vormgevers, Ede, The Netherlands

Text Mining to Support Knowledge Discovery from Electronic Health Records

Tekstmining als hulp bij het ontdekken van kennis uit elektronische patiëntendossiers

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties
De openbare verdediging zal plaatsvinden op

dinsdag 3 juli 2018 om 11.30 uur
door

Muhammad Zubair Afzal
geboren te Rawalpindi, Pakistan

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:

Prof.dr. M.C.J.M. Sturkenboom

Overige leden:

Prof.dr. J.A. Hazelzet

Prof.dr. A. van den Bosch

Prof.dr. A. Abu-Hannah

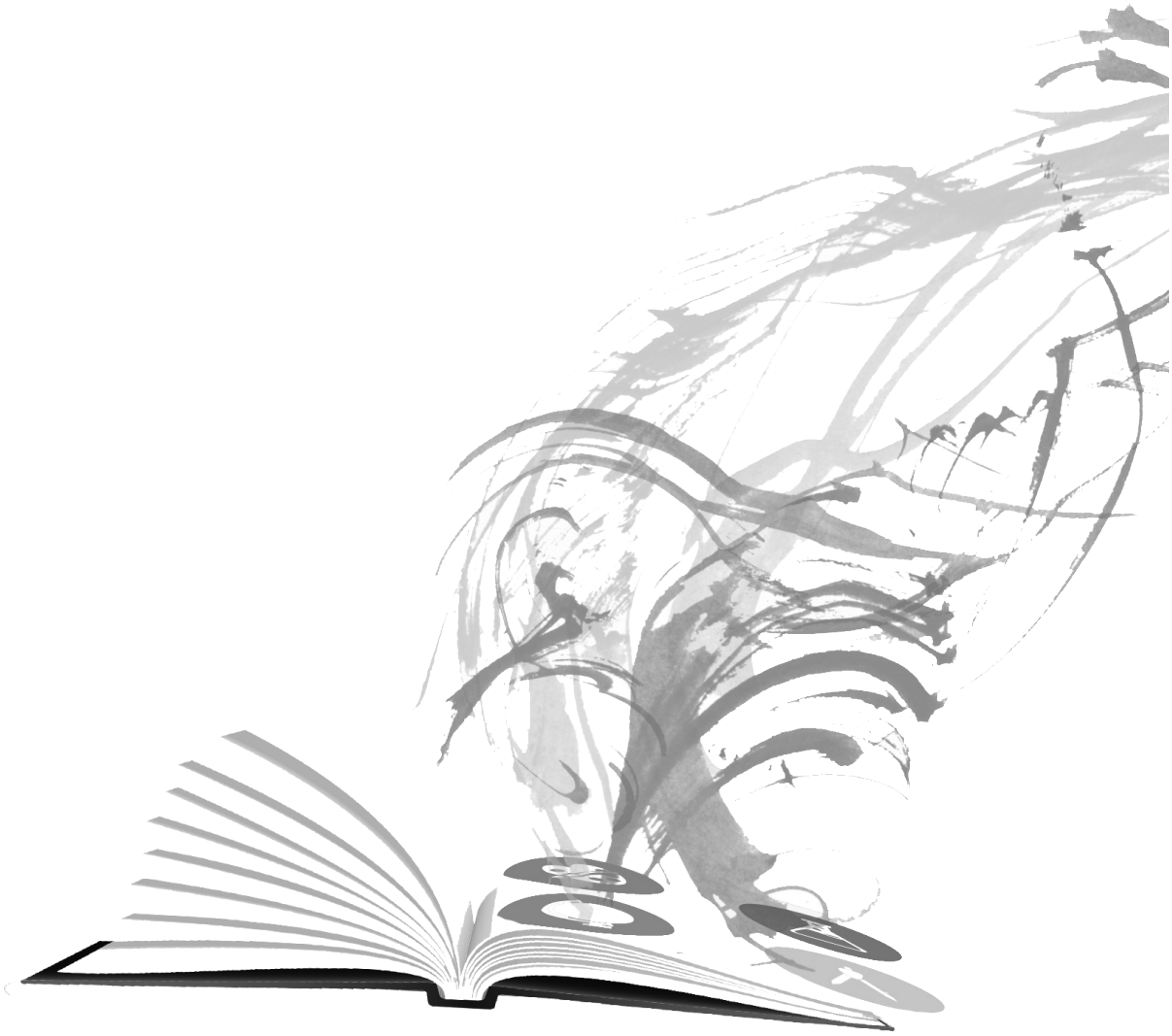
Copromotoren:

Dr.ir. J.A. Kors

Dr. M.J. Schuemie

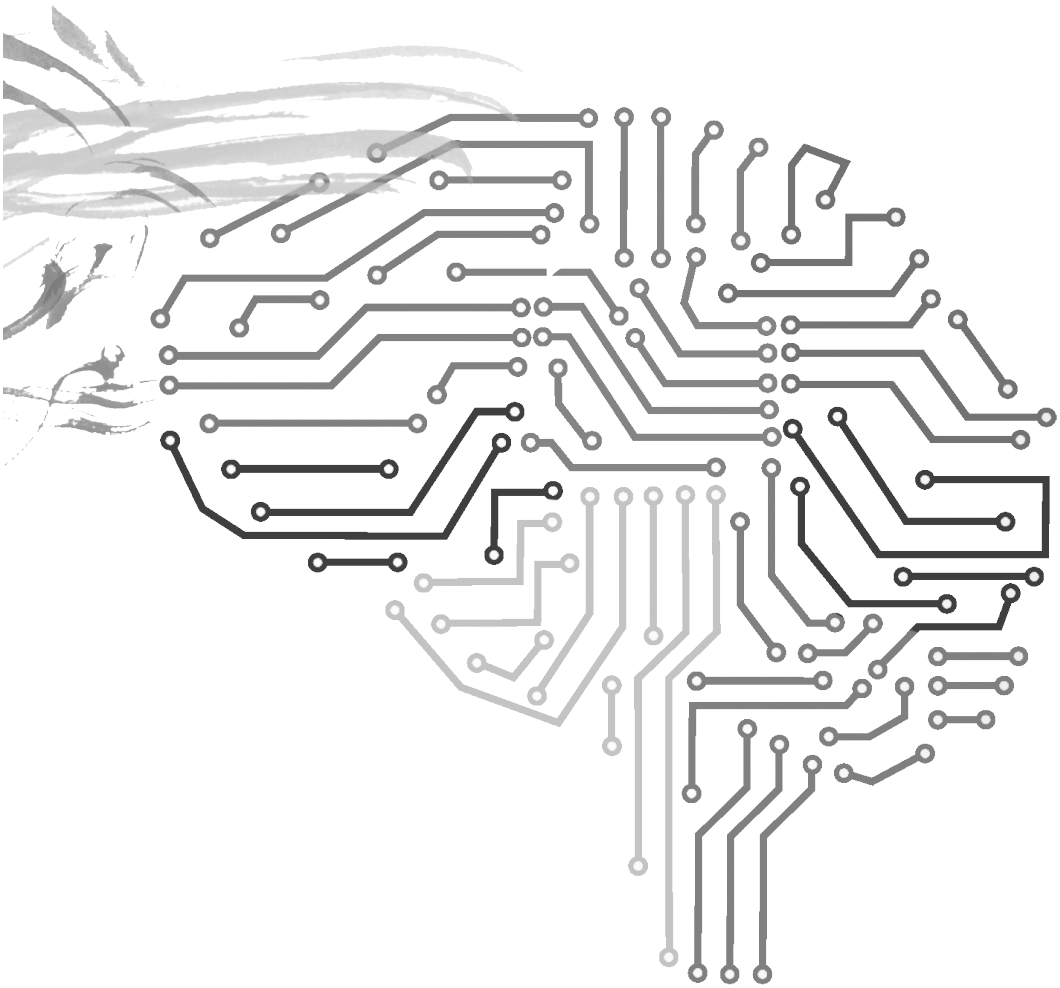
For My Family

TABLE OF CONTENTS



Chapter 1

Introduction



-

The use of electronic health records (EHRs) has grown rapidly in the last decade. The EHRs are no longer being used only for storing information for clinical purposes but the secondary use of the data in the healthcare research has increased rapidly as well. The data in EHRs are recorded in a structured manner as much as possible, however, many EHRs often also contain large amount of unstructured free-text. The structured and unstructured clinical data presents several challenges to the researchers since the data are not primarily collected for research purposes. The issues related to structured data can be missing data, noise, and inconsistency. The unstructured free-text is even more challenging to use since they often have no fixed format and may vary from clinician to clinician and from database to database. Text and data mining techniques are increasingly being used to effectively and efficiently process large EHRs for research purposes. Most of the methods developed for this purpose deal with English-language EHRs and cannot simply be applied to non-English EHRs. This thesis concerns the use of data mining and natural language processing techniques to process unstructured Dutch-language EHRs. We present all methods and approaches in this thesis in a wider and formal framework of knowledge discovery. We begin with an introduction to knowledge discovery, and subsequently we will continue by describing a knowledge discovery pipeline. Following that, we will describe general data preparation and data mining techniques in more detail. Further, we will present electronic health records and its challenges. Finally, we end with an outline of the work done in this thesis.

Knowledge discovery

Recent years have seen an exponential increase in the generation and collection of all sorts of data for various purposes. It is only natural that such growth in data generation and collection is also matched with a growing number of methods and techniques for knowledge discovery. The process of knowledge discovery consists of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. The exact definitions of valid, novel, useful, and understandable information depends on the actual knowledge discovery task. The term data mining is often used interchangeably with knowledge discovery but in fact, data mining is a part of the knowledge-discovery process. We will first look at general knowledge-discovery models and then we will shift our focus onto data preparation and data mining.

Over the last decades, several generic process models have been proposed for knowledge discovery. The most common or popular ones include the nine-step model proposed by Fayyad et al. [2], the industry validated six-step CRISP-DM (CRoss Industry Standard Process for Data Mining) [3] developed by a consortium of European companies, the five-step model proposed by Cabena et al. [4], and the six-step DMKD (Data Mining for Knowledge Discovery) model proposed by Cios et al. [5]. All these models share steps of first understanding the problem and the data, building methods for knowledge extraction, and evaluating the extracted knowledge. The DMKD model (Figure 1) is actually a modified and improved version of the CRISP-DM model and the model proposed by Cabena et al. One of the main differences is that the DMKD model includes several new feedback mechanisms as compared to only three feedbacks in the CRISP-DM model. The feedback mechanisms are represented with dotted lines in Figure 1. A detailed comparison of several knowledge discovery models is presented in [6].

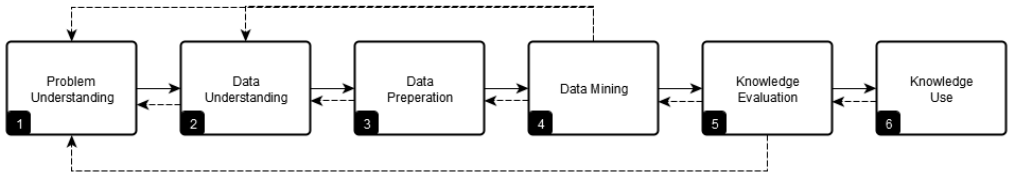


Figure 1: Data Mining for Knowledge Discovery (DMKD) process

Problem Understanding

The first step in the process is about understanding the problem from multiple perspectives such as domain-specific requirements, defining the objectives, goals, and success criteria, understanding data privacy issues (especially for clinical data), engaging domain experts, and exploration of existing solutions that could lead to an initial selection of tools and methods to be used during the data-mining step.

Data Understanding

The next important step in the process is about understanding the data that will be used for the data-mining process. This involves understanding the format of the data in which it is available, such as structured tables, databases, or unstructured free-text, and if the data would be sufficient to answer the questions raised in the first step. Typically, a sample of the data is collected and checked for completeness and quality.

Data Preparation

This is one of the crucial steps and perhaps the most time-consuming in the process. In this step, it is decided which and how much of the data will be used for data mining and then the data are collected for further processing. The sub-tasks depend on the current state of the data and the questions one is trying to answer through the knowledge-discovery process. Typically, it involves cleaning, normalizing, sampling, reducing dimensionality, summarizing and aggregating, extracting features, and transforming the data to a format that is expected by the data-mining tools.

Data Mining

In this step, different data-mining tools or algorithms are used on the prepared data to extract information or derive knowledge. The selection of data-mining method(s) is a crucial step and often multiple methods are selected initially based on previous experience, the type of data available, and the objectives to achieve. There is a large pool of such methods to choose from such as neural networks, Bayesian methods, support vector machines (SVM), decision trees, rule learners, and clustering. Many of these algorithms allow tweaking of algorithm parameters to

achieve better results. Špečkauskienė et al. [7] proposed a very detailed and iterative 11-step process to select the optimum data-mining method. However, for very large datasets, this 11-step selection process may prove to be too time-consuming and resource intensive and in some cases it may just not be feasible at all [8].

Knowledge Evaluation

The next step in the knowledge-discovery process is to understand the results and analyze the information that is extracted from the data. Usually results and the extracted information are evaluated together with the domain experts to fully understand the usefulness and the novelty. The objectives and goals are revisited to decide if further iterations are required to improve the results.

Knowledge Use

The last step in the process is to plan how the discovered knowledge will be used. This could involve turning the experimental setup to a product or a service so the process can be repeated for new yet similar data sets easily. The experimental setup with details of all decisions and assumptions is also documented in this step.

In the next section, we will focus more on the two most important steps in the knowledge-discovery process, i.e., data preparation and data mining.

Data Preparation

As mentioned earlier, data preparation is one of the most crucial and most time-consuming steps in the knowledge discovery process. It is estimated that about 60% to 90% of the time is spent on this step [5, 9]. The data collected in real-world applications are not always directly suitable for data-mining tasks. Almost all real-word data sets suffer from inconsistency, redundancy, incompleteness, contradictions, and noise to some extent. These issues may have strong impact on the entire outcome of the knowledge-discovery process. Therefore, it is critical that the data are prepared carefully for subsequent data-mining task. The tasks in this step depend on the current state of the data, the objectives of the knowledge-discovery process, and the algorithms and tools to use in the subsequent processes. In a broader sense, the sub-tasks in the data preparation stage can be categorized into the following groups:

1. Data Cleaning

The data set might be missing some important attributes of interest or missing some attribute values, e.g., missing blood pressure measures for some patient. There are typically two ways to deal with the missing data: a) ignoring the data point with missing values completely, or b) fill in the missing values such as using NULL or using some advanced techniques to infer the most probable value to fill. It is also important to look for inconsistencies in the data such as the use of different date formats and use of different disease classification schemes in different years. Such inconsistencies must be dealt with properly before data are used. Another important data-

cleaning technique deals with smoothing out the noisy data. The smoothing may include identifying and correcting, disambiguating, removing, or ignoring noise in the data set, e.g., correcting spelling mistakes, grouping similar words together, correcting negative values where only positive values are expected, and identifying outliers.

2. Data Integration

Data integration deals with properly integrating or combining different sources of data into one data set for further processing. It is possible that the data are recorded differently in different sources, e.g., in one data set, patient diagnoses are coded using International Classification of Diseases (ICD) version 9, and ICD version 10 is used in another data set. This needs to be corrected to avoid data inconsistencies. During integration, redundant data attributes are also identified and removed.

3. Data Transformation

For distance-based data mining methods such as nearest neighbors, it is important that all attributes have the same units and scales to allow a fair comparison. Different normalization techniques can be used to transform different types of data, e.g., min-max normalization can scale all the numeric values between 0 and 1. Another important transformation technique in the data preparation phase is called discretization. Some data-mining algorithms are not capable of handling continuous feature values, therefore, features with continuous values needs to be converted to nominal or categorical features. For example, systolic blood pressure of 140 and above can be categorized as ‘high’ and less than 80 can be categorized as ‘low’.

4. Data Extraction

Data extraction techniques deal with processing and extracting important and relevant information. All data-related inclusion and exclusion criteria are enforced to meet the study requirements. Several techniques can be used to extract features for the subsequent data-mining tasks. For data sets with free-text, feature extraction typically involves natural language processing (NLP) techniques. NLP is a specialized field in artificial intelligence that deals with applying computational techniques for the analysis and synthesis of natural language and speech data. The term NLP is often used synonymously with text mining. Several NLP techniques such as sentence splitting, tokenization, part-of-speech tagging, chunking, named entity recognition, and relation extraction can be used to extract relevant features from unstructured free-text [10, 11]. In sentence splitting, paragraphs of text are divided into single sentences. The sentences are further split into words and symbols called tokens. Part-of-speech (POS) tagging is the process of assigning correct POS tag such as noun and verb to each word in a sentence. Chunking is the process of using POS tags to identify phrases within a sentence such as noun phrase, verb phrase, or prepositional phrase. Name entity recognition (NER) is the process of identifying predefined entity types in the text such as personal names, locations, drugs, diseases, and symptoms.

5. Data Reduction

Data reduction techniques focus on removing irrelevant features or data elements. Several techniques such as data aggregation, dimensionality reduction, and clustering can be used to reduce the data set effectively. In data aggregation, data is reduced to represent a higher level of information, e.g., months can be aggregated into quarters or city names to state names. In dimensionality reduction, all irrelevant features and noisy data is removed. This leads to a reduction in time and space required for machine-learning algorithms. All words with the same meaning can be combined into one using clustering techniques to reduce features even further.

Data Mining

Data mining is the process of analyzing large datasets to identify and extract hidden patterns or information. Data-mining methods are usually computationally intensive and yield patterns that are valid, novel, useful, and human understandable. This is typically achieved by employing statistical or machine-learning techniques. Machine learning is formally defined by Tom Mitchell [12] as: 'A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*'. For example, a task *T* could be to identify patients with a disease of interest from a large database. The experience *E* is a known set of patients with the disease and the performance *P* is the accuracy of the algorithm to identify patients from unseen data. Some other common data-mining task examples are identifying patients with various conditions, identifying candidates in need of therapy, spam detection, fraud detection, loan/credit approval, identifying customers who are likely to cancel their subscriptions, finding relationships between diseases, predicting sales forecast, etc.

Data-mining approaches can be broadly categorized into supervised or predictive learning and unsupervised or descriptive learning methods. In supervised learning methods, the algorithm is usually provided with some examples where the output or the class labels are already known. Such examples constitute a training set. The algorithm utilizes the training set to learn or build a model that can later predict the output of the unseen data, which constitute the test set. For example, in a spam detection task, the algorithm is given a set of emails (i.e., training set) where each email is labeled as either a spam email or not. The algorithm then learns a model (also called a classifier) that can later classify if an incoming email is spam or not. In unsupervised learning methods, the provided training set does not contain the known desired output. Such methods try to describe the data by learning underlying structure or patterns such as grouping similar-looking items.

The next sections describe classification and clustering, two examples of supervised and unsupervised learning, respectively.

Classification

Classification is the task of predicting group membership as an output for new data instances. Since it is an example of a supervised learning method, a classification algorithm (also known as

a classifier) first learns how to predict using the training set where the desired output is already known. There are two types of classifier prediction output. The first type is where the classifier assigns a class label to new instances. The set of class labels to choose from is predefined in this case. If the classifier has to choose between two class labels (e.g., Yes or No), it is called binary classification. If there are more than two class labels to choose from, it is called multi-class classification. There is another category where the classifier has to choose from more than two labels but also multiple labels can be selected for each instance. This is called multi-class multi-label classification, which is more challenging. In the second type of classifier prediction output, for each new instance, the classifier returns a probability (of belongingness) for each class label. For example, what is the probability for a patient to have disease A or B? This is also known as probabilistic classification.

Generally, the data set is divided into two subsets, a training set and a test set. The algorithms first use the training set for learning a classifier and then use the test set to estimate the actual classification performance. Sometimes, the training data are further divided in two sets where one is used for learning the classifier and the second to estimate the classification error and tune the parameters. The purpose of this further division of the training set is to avoid the *overfitting* problem. Overfitting is when the classifier performs well on the training set but performs poorly on the test set. It means that the classifier has learned the data but not the actual underlying function.

Clustering

Clustering is one of the most frequently used unsupervised learning approaches. Since the desired output is not known for the training examples, this method relies on finding the similarities and differences between the examples in order to create subsets, called clusters. The purpose is to create clusters where all members in the same cluster are similar. Grouping similar customers together for target advertising is a typical example of a clustering application [13, 14]. The ideal clustering is when members of the same cluster have homogeneity and the clusters are very different from each other. There are three popular approaches to clustering [14, 15]. In the first approach, new clusters are derived from existing clusters. This is also known as hierarchical clustering. In this approach, new clusters are derived either by merging similar clusters together or by splitting existing clusters into smaller ones. In the second clustering approach, first, a pre-specified number of clusters are created and then cluster membership is improved by moving cluster members from one cluster to another. In the third clustering approach, instances with close vicinity are combined into one cluster. This is also known as density-based clustering.

Electronic Health Records

The primary use of electronic health records (EHRs) is to support the care process by the clinicians and the administrators [16]. These records contain patient-related information such as demographics (e.g., age, sex, ethnicity), behavior (e.g. use of tobacco and alcohol), vital signs (e.g., body temperature and blood pressure), patient-reported symptoms (e.g., headache), diagnosis (e.g., hypertension), procedures (e.g., electrocardiogram), treatments (e.g.,

medications and their usage information), laboratory data (e.g., test reports), allergies, and imaging (e.g., CT scans).

The data in EHRs are recorded in a structured manner as much as possible. Often structured clinical terminologies are used such as ICD-10, Logical Observational Identifiers Names and Codes (LOINC), and Anatomical Therapeutic Chemical (ATC) codes to record information. Most EHR systems use a predefined way of storing structured information (e.g., using schemas) in order to avoid recording errors and variations. This allows easy and quick access to the structured data for reporting and analytical purposes.

However, storing all information in a structured way may not be practical, as it requires knowing exactly what to expect and store beforehand. This may lead to an increase in the complexity of the EHR systems and decrease their usability [17]. All EHR systems allow healthcare providers to also record information in free-text, which is unstructured and gives clinicians maximum flexibility to record anything regarding the patient. The information present in the free-text often contains essential information such as patient-reported symptoms, signs, summaries of specialists' letters in narrative form, past medical history, family medical history, behavior and lifestyle information. This information may be critical for identification of the medical events. Since there is no standard way of writing clinical narratives, the style of writing, the amount of information, and the use of language may vary from clinician to clinician. Some clinical documents such as discharge letters or referrals are used as a formal way of communication between clinicians. Other documents may be used as references such as nurses' daily notes. The clinical notes are usually written under time pressure and they often contain ill-formed and incomplete sentences. Apart from that, there are also grammatical errors, standard and non-standard abbreviations, and misspellings. Ruch et al. [18] reported up to 10% spelling errors in follow-up nursing notes. The information recorded in free-text may also be inconsistent. Wasserman et al. [19] found 278 different ways of reporting fever in clinical notes of 465 children. Extracting useful information from unstructured free-text is a challenging task, which usually requires advanced Natural Language Processing (NLP) techniques [20].

Although the primary use of EHRs is to support the care process, the secondary use of EHRs for clinical research is increasing. However, this also poses several other challenges apart from the ones mentioned above regarding the free-text. For example, in the absence of a nationwide healthcare system, if a patient has been seen by multiple healthcare providers (such as different specialists) over the years, then all providers may only have partial information about the patient. This is known as information fragmentation [17]. Linking information from different providers is not trivial as they might be using different EHR system providers or, even if they are using the same EHR system provider, the way the information is recorded might be different due to local policies. Structured codes such as ICD-10 used by the care providers may not be accurate or at times missing [21].

Mining Electronic Health Records

EHRs enable researchers to utilize rich longitudinal health data at a relatively low cost. Collecting such large amounts of data otherwise would be expensive and time-consuming. Data mining

techniques are increasingly being used in observational epidemiological studies that are utilizing EHRs, for example, in studies that aim to investigate the association between drugs and possible adverse events [22–24]. An important initial first step in such studies is the identification of the patients who have the event of interest, commonly known as case selection. The traditional approach involves issuing a broad query to first identify potential patients and then manually reviewing patient data to distinguish true positive cases from true negative cases. This process is also known as manual chart review. Manual review is expensive, time consuming and becoming prohibitive with the increasing size of EHR databases. Automated methods are being used to identify patients with various conditions [25–29], candidates in need of therapy [30], smoking status [31, 32], patients who are similar to a patient under observation [33], direct and indirect associations among medical concepts such as diseases and medications [34], events related to adverse drug-drug associations [35], and to analyze medications and food allergies [36]. One particular challenge in analyzing free-text EHRs is to distinguish positive diagnoses from things that have been excluded. For example, in the text '*the patient was diagnosed with asthma',* the patient is positively diagnosed with asthma, however, in the text '*the patient was not diagnosed with asthma'*, the clinician is explicitly ruling out the presence of asthma. Several machine-learning approaches have been used to identify negated concepts and their scopes from various types of clinical documents [37–40]. Yadav et al. [17] presented an extensive survey of various data-mining techniques that are used to model EHR data.

Thesis Outline

Despite the increasing use of data-mining techniques in the healthcare domain, it is still considered to be in early stages as compared to other domains [17]. Most of the observational studies in the field use structured information present in the EHRs. Only a few studies have tried to exploit the information present in the unstructured free-text. In addition, the methods employed or developed to process and extract information from free-text primarily have focused on English EHRs. This thesis aims at developing automated methods to exploit unstructured free-text present in the Integrated Primary Care Information (IPCI) database [41], a longitudinal collection of EHRs from Dutch general practitioners.

In order to extract meaningful information from EHRs, such as symptoms and diagnoses, it is also important to identify their contextual properties. Whether a diagnosis is positive or negative, a medical condition described in the patient record is new or old, or is about the patient or someone else (e.g., a family member) are important questions that require understanding of the context in which information is provided. In Chapter 2, we adapt an English language algorithm to the Dutch language in order to identify contextual properties of clinical concepts. A new Dutch clinical corpus has been created to evaluate the performance of our system.

In Chapter 3, we look at the methods to normalize free-text in Dutch EHRs. The normalization process involves grouping similar words together to reduce feature dimensionality for machine-learning methods and automatically finding and mapping abbreviations and acronyms to their full-forms in the database.

Observational studies need to deal with confounding in order to obtain unbiased estimates. Confounding occurs when a variable (i.e., confounder) that is not under investigation, influences the outcome of interest. EHRs contain much unstructured data that could be used as proxies for potential confounding factors. In Chapter 4, we look at the possibility of using unstructured free-text to construct high-dimensional propensity score models that would allow to properly deal with confounding.

Case selection is one of most important and time-consuming tasks in observational studies. In Chapter 5, we use and evaluate machine-learning methods to automatically generate case-detection algorithms for case selection that uses both free-text and structured information in the EHRs.

Typically, the proportions of positive and negative cases in the training set are not equal as there are usually more negative cases than positive cases. Such imbalance affects the learning process of the supervised methods. Finally, in Chapter 6, we use different approaches to handle imbalance in the training set in order to improve the performance of case-detection algorithms.

Table 1: Overview of the topics described in this thesis

Chapter	Research topic	Data used	Knowledge discovery step
2	ContextD: An algorithm to identify contextual properties of medical terms in a Dutch clinical corpus	IPCI, DL, RD	Data preparation
3	Reducing feature dimensionality by normalizing text in electronic health records	IPCI	Data preparation
4	Generating and evaluating a propensity model using textual features from electronic health records	IPCI	Data preparation
5	Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases	IPCI	Data mining
6	Improving sensitivity of machine learning methods for automated case identification from free-text medical records	IPCI	Data mining

IPCI, Integrated Primary Care Information; DL, Discharge Letters; RD, Radiology Reports

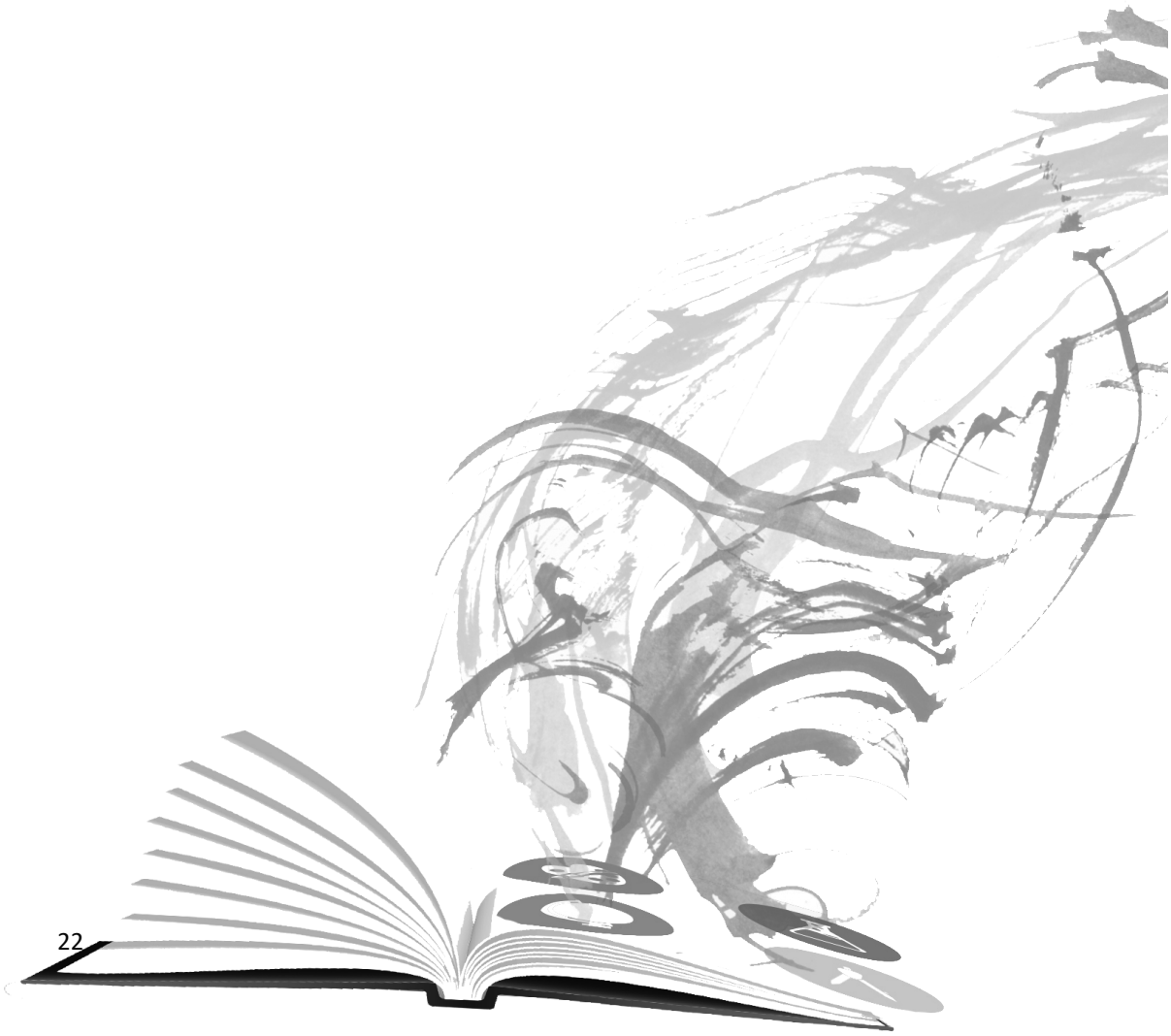
REFERENCES

1. Fayyad UM, Piatetsky-Shapiro G, Smyth P: **From data mining to knowledge discovery: An overview**. In *Adv Knowl Discov Data Min*. Edited by Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Menlo Park, CA, USA: American Association for Artificial Intelligence; 1996:1–34.
2. Fayyad UM, Piatetsky-Shapiro G, Smyth P: **Knowledge discovery and data mining: Towards a unifying framework.** In *Proc Second Int Conf Knowl Discov Data Min*. AAAI Press; 1996:82–88.
3. Wirth R: **CRISP-DM : Towards a standard process model for data mining**. In *Proc Fourth Int Conf Pract Appl Knowl Discov Data Min*; 2000:29–39.
4. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A: *Discovering data mining: From concept to implementation*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; 1998.
5. Cios K, Kurgan L: *Trends in data mining and knowledge discovery*. Springer London; 2005(Dm).
6. Klösgen W, Zytkow JM: *The knowledge discovery process*. New York, NY, USA: Oxford University Press, Inc.; 2002.
7. Špečkauskienė V, Lukoševičius A: **Methodology of adaptation of data mining methods for medical decision support : Case study**. *Electron Electr Eng* 2009, **2**:25–28.
8. Niakšu O: **CRISP data mining methodology extension for medical domain**. *Balt J Mod Comput* 2015, **3**:92–109.
9. YANG Q, WU X: **10 Challenging problems in data mining research**. *Int J Inf Technol Decis Mak* 2006, **5**:597–604.
10. Nadkarni PM, Ohno-Machado L, Chapman WW: **Natural Language Processing: An introduction**. *J Am Med Informatics Assoc* 2011, **18**:544–551.
11. Indurkhya N, Damerau FJ: *Handbook of Natural Language Processing*. *Volume 2*. CRC Press; 2010.
12. Mitchell TM: *Machine Learning*. *Volume 4*. McGraw-Hill; 1997. [*McGraw-Hill Series in Computer Science*]
13. Coenen F: **Data Mining : Past , present and future**. *Knowl Eng Rev* 2004:25–29.
14. Mirkin B: *Clustering for data mining - A data recovery approach*. 2005.
15. Jain AK: **Data clustering: 50 years beyond K-means**. *Pattern Recognit Lett* 2010, **31**:651–666.
16. Casey JA, Schwartz BS, Stewart WF, Adler NE: **Using electronic health records for population health research: A review of methods and applications**. *Annu Rev Public Health* 2016, **37**:61–81.
17. Yadav P, Steinbach M, Kumar V, Simon G: **Mining electronic health records: A survey**. *ACM Comput Surv* 2016, **1**:1–41.
18. Ruch P, Baud R, Geissbühler A: **Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record.** *Artif Intell Med* 2003, **29**:169–184.
19. Wasserman RC: **Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research.** *Acad Pediatr* 2011, **11**:280–287.

20. Demner-Fushman D, Chapman WW, McDonald CJ: **What can natural language processing do for clinical decision support?** *J Biomed Inform* 2009, **42**:760–772.
21. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM: **Measuring diagnoses: ICD code accuracy**. *Health Serv Res* 2005(5 II):1620–1639.
22. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW: **Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting**. *Pharmacoepidemiol Drug Saf* 2010, **19**:1211–1215.
23. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR: **Temporal pattern discovery in longitudinal electronic patient records**. *Data Min Knowl Discov* 2009, **20**:361–387.
24. Boockvar KS, Livote EE, Goldstein N, Nebeker JR, Siu A, Fried T: **Electronic health records and adverse drug events after patient transfer.** *Qual Saf Health Care* 2010, **19**:e16.
25. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM: **Electronic medical records for discovery research in rheumatoid arthritis.** *Arthritis Care Res (Hoboken)* 2010, **62**:1120–1127.
26. Perlis RH, Iosifescu D V, Castro VM, Murphy SN, Gainer VS, Minnier J, Cai T, Goryachev S, Zeng Q, Gallagher PJ, Fava M, Weilburg JB, Churchill SE, Kohane IS, Smoller JW: **Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model.** *Psychol Med* 2012, **42**:41–50.
27. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH: **NLP-based identification of pneumonia cases from free-text radiological reports.** In *AMIA Annu Symp Proc*; 2008:172–176.
28. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, Kullo IJ: **Discovering peripheral arterial disease cases from radiology notes using natural language processing.** In *AMIA Annu Symp Proc*. *Volume 2010*; 2010:722–726.
29. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL: **Electronic medical records for clinical research: application to the identification of heart failure.** *Am J Manag Care* 2007, **13**(6 Part 1):281–288.
30. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW: **Electronic health record-based cardiac risk assessment and identification of unmet preventive needs.** *Med Care* 2009, **47**:418–424.
31. Savova GK, Ogren P V, Duffy PH, Buntrock JD, Chute CG: **Mayo clinic NLP system for patient smoking status identification.** *J Am Med Inform Assoc* 2008, **15**:25–28.
32. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U: **Identifying smokers with a medical extraction system.** *J Am Med Inform Assoc* 2007, **15**:36–39.
33. Gotz D, Sun J, Cao N, Ebadollahi S: **Visual cluster analysis in support of clinical decision intelligence.** *AMIA Annu Symp Proc* 2011, **2011**:481–490.
34. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G: **Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics.** *AMIA Annu Symp Proc* 2005:106–110.
35. Iyer S V, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH: **Mining clinical text for signals of adverse drug-drug interactions**. *J Am Med Informatics Assoc* 2014, **21**:353–362.
36. Epstein RH, St Jacques P, Stockin M, Rothman B, Ehrenfeld JM, Denny JC: **Automated identification of drug and food allergies entered using non-standard terminology**. *J Am*

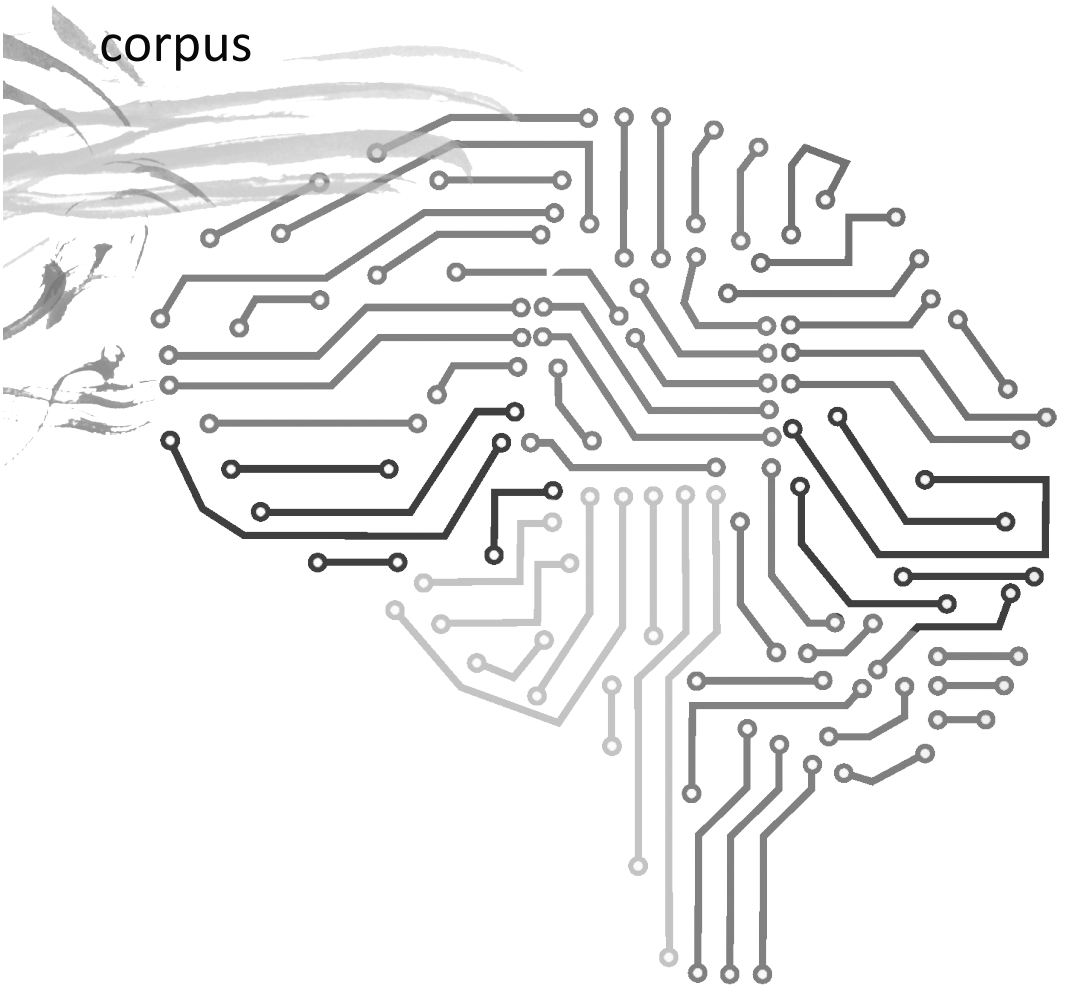
Med Informatics Assoc 2013, **20**:962–968.

37. Goldin I, Chapman W: **Learning to detect negation with "not" in medical texts**. In *Proc ACM-SIGIR 2003*; 2003.
38. Agarwal S, Yu H: **Biomedical negation scope detection with conditional random fields.** *J Am Med Inform Assoc* 2010, **17**:696–701.
39. Morante R, Daelemans W: **A metalearning approach to processing the scope of negation**. In *Proc Thirteen Conf Comput Nat Lang Learn - CoNLL '09*. Boulder, Colorado; 2009(June):21–29.
40. Cruz Díaz NP, Maña López MJ, Vázquez JM, Álvarez VP: **A machine-learning approach to negation and speculation detection in clinical texts**. *J Am Soc Inf Sci Technol* 2012, **63**:1398–1410.
41. Vlug A, van der Lei J, Mosseveld B, van Wijk M, van der Linden P, MC S, van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project**. *Methods Inf Med* 1999, **38**:339–344.



Chapter 2

ContextD: An algorithm to identify contextual properties of medical terms in Dutch clinical corpus



ABSTRACT

Background

In order to extract meaningful information from electronic medical records, such as signs and symptoms, diagnoses, and treatments, it is important to take into account the contextual properties of the identified information: negation, temporality, and experiencer. Most work on automatic identification of these contextual properties has been done on English clinical text. This study presents ContextD, an adaptation of the English ConText algorithm to the Dutch language, and a Dutch clinical corpus. We created a Dutch clinical corpus containing four types of anonymized clinical documents: entries from general practitioners, specialists' letters, radiology reports, and discharge letters. Using a Dutch list of medical terms extracted from the Unified Medical Language System, we identified medical terms in the corpus with exact matching. The identified terms were annotated for negation, temporality, and experiencer properties. To adapt the ConText algorithm, we translated English trigger terms to Dutch and added several general and document specific enhancements, such as negation rules for general practitioners' entries and a regular expression based temporality module.

Results

The ContextD algorithm utilized 41 unique triggers to identify the contextual properties in the clinical corpus. For the negation property, the algorithm obtained an F-score from 87% to 93% for the different document types. For the experiencer property, the F-score was 99% to 100%. For the historical and hypothetical values of the temporality property, F-scores ranged from 26% to 54% and from 13% to 44%, respectively.

Conclusions

The ContextD showed good performance in identifying negation and experiencer property values across all Dutch clinical document types. Accurate identification of the temporality property proved to be difficult and requires further work. The anonymized and annotated Dutch clinical corpus can serve as a useful resource for further algorithm development.

BACKGROUND

Recent years have seen an increase in the use of electronic medical records (EMRs) by healthcare providers [1]. These records contain patient-related information such as signs, (patient-reported) symptoms, diagnoses, treatments, and tests. The primary use of EMRs is to support the care process, but the secondary use of EMRs for clinical research is increasing. In most EMRs, the majority of information is unstructured free text, making information retrieval challenging, although several automatic systems have been developed that can index, extract, and encode clinical information from the EMRs [2-8]. One particular challenge in analyzing free-text EMRs is to distinguish positive diagnoses by the physician from things that have been excluded or ruled out. Similarly, information about the past medical problems and a family history is often found in the EMRs and should ideally be identified as such. In order to extract meaningful information such as medical problems or clinical conditions, it is important that automatic systems do not only identify them but also take into account the context of the identified information.

Previous approaches on identifying contextual properties of clinical concepts can be classified into rule or regular-expression based techniques, machine-learning techniques, or a combination of both. Chapman et al. [9] developed a rule-based system called NegEx that determines whether a specific medical condition is present or absent within a narrative. The system uses two sets of trigger phrases: one to identify true negations and a second to identify pseudo-negations, i.e., phrases that seem to indicate negation but instead denote double negations such as not ruled out. The system was evaluated on dis- charge summaries where it achieved a precision of 84.5% and a recall of 77.8%. Another system, called NegFinder [10], used grammatical parsing and regular expressions to identify negated patterns occurring in medical narratives, achieving a specificity of 97.7% and a sensitivity (or recall) of 95.3% on discharge summaries and surgical notes. Elkin et al. [11] assigned a level of certainty to identified concepts in EMRs based on a rule-based system to decide whether a concept has been asserted positively, negatively, or uncertainly. The system achieved 97.2% sensitivity, 98.8% specificity, and 91.2% precision on medical evaluation notes. Huang et al. [12] used regular expressions with grammatical parsing to identify ne- gated phrases. On radiology reports, the system achieved a sensitivity of 92.6%, a specificity of 99.8%, and a precision of 98.6%. The ConText algorithm [13] is based on the NegEx algorithm and apart from identifying negations; it identifies whether a clinical condition is present, historical, or hypothetical, and whether the patient or someone else, e.g., a family member, experiences the clinical condition. The system achieved an average precision of 94% and an average recall of 92% when evaluated on six different types of medical reports. Kilicoglu and Bergler [14] showed that speculative language can be recognized successfully using linguistically oriented approaches. They extended lexical resources with syntactic patterns and introduced a simple weighting scheme to estimate the speculation level of the sentences. The system achieved a precision of 85% and a recall of 86%. Recently, Reeves et al. [15] created a system, Med-TTK, to identify and classify temporal expressions in medical narratives. The system achieved a precision of 85% and a recall of 86% on clinical notes.

In machine-learning approaches, Goldin and Chapman [16] experimented with Naïve Bayes and decision trees to determine whether a concept is negated by the word not in hospital progress notes and emergency room notes. Agarwal and Yu [17] used conditional random fields (CRF) to detect negation cues and their scopes. The best CRF model achieved a precision of 99% and a recall of 96% on detecting negation cues, and a precision and recall of 95% on detecting their scopes in clinical notes. Morante and Daelemans [18] first used a classifier to identify negation signals and then used four classifiers to find the full scope of the negation signals. Three of the classifiers predicted whether a token was the first, the last, or neither in the scope sequence. The fourth classifier was a meta-learner that used the prediction of first three classifiers to determine the final scope. On BioScope clinical documents [19], the system achieved a precision of 86% and a recall of 82%, and 71% of negation scopes were correctly identified. Cruz Díaz et al. [20] improved on Morante and Daelemans [18] by using different classifiers. The system achieved a precision of 92%, a recall of 90%, and 88% of the negation scopes were correctly identified. To detect speculation, the system achieved a precision of 85%, a recall of 63%, and 63% of speculation scopes were correctly identified. Light et al. [21] estimated that 11% of the sentences in MEDLINE abstracts contain speculative fragments. They used a substring matching method and Support Vector Machines (SVM) to determine whether concepts in the text are described as facts or as speculation. For the matching method, they identified 14 strings that suggest speculation and marked a sentence as speculative if their system found any of these strings in the sentence (possibly as a substring of a term). The SVM classifier achieved a precision of 84% and a recall of 37%, whereas substring matching achieved a precision of 55% and a recall of 79%. Velldal [22] used a disambiguation approach and SVM-based classifiers to label sentences as certain or uncertain. Their best system achieved a precision of 89% and a recall of 85%. Goryachev et al. [23] compared two adaptations of regular-expression based algorithms, NegEx and NegExpander, with two classification methods, Naïve Bayes and SVM, trained on dis- charge reports. It was observed that regular-expression based methods show better accuracy than the classification methods. Uzuner et al. [24] developed a statistical assertion classifier, StAC, by using lexical and syntactic context in conjunction with SVM to classify medical problems in EMRs into four categories: positive, negative, uncertain, and alter-association. StAC was compared to an extended version of the NegEx algorithm and showed better performance. The 2012 i2b2 NLP Shared Task [25] focused on finding the temporal relations in clinical narratives. While machine-learning and rule-based systems showed good performance, the systems using combination approaches produced the best results.

The type of clinical documents has a noticeable impact on the performance of systems that identify contextual properties of clinical concepts. Clinical documents differ in many ways, such as structure, grammaticality, and use of standard and non-standard abbreviations. Overall, there does not seem to be a clear winner between machine- learning and rule-based systems. The rule-based and hybrid systems appear to perform slightly better than machine-learning approaches. In theory, rule-based systems can be adapted rather easily for different clinical text than for which they were developed. One of the limiting factors of a rule-based approach is the use of a fixed scope, which may lead to misclassification. The machine-learning based approaches may not perform as well if they are tested on a different clinical text than they were originally trained on [23]. Adapting such approaches for new clinical text will therefore require a new training set.

Most work on identifying contextual properties of the clinical condition has been done on the English language. Recently, the NegEx algorithm was adapted to detect negations in Swedish [26] and French [27] clinical text. To our knowledge, no method is yet available or adapted for Dutch clinical text.

This study has two objectives: to adapt the well-known ConText [13] algorithm (to detect contextual properties of medical terms) to the Dutch language and to create a Dutch clinical corpus which is annotated for negation, temporality, and experiencer. ConText, along with its predecessor NegEx, is one of the most widely used algorithms in the field. It was chosen for its simplicity, ease of adaptability, and proven good performance on various types of English clinical text. The adapted ConText algorithm, dubbed ContextD, and the anonymized Dutch clinical corpus described here will be made publicly available for research purposes [28].

METHODS

This section provides details of the Erasmus Medical Center (EMC) Dutch clinical corpus annotated for the three contextual properties negation, temporality, and experiencer. We also describe the original ConText algorithm and its adaptation to the Dutch language.

EMC Dutch clinical corpus

The anonymized corpus includes four types of clinical documents to capture different language use in the Dutch clinical setting.

- *General Practitioner entries [GP]*

  This set consists of entries from the IPCI database [29], a longitudinal collection of EMRs from Dutch general practitioners (GP) covering more than 1.5 million patients throughout the Netherlands. Each entry in the IPCI database pertains to a patient visit to the GP. These entries are not always grammatically well-formed text, and often follow the well-known SOAP structure (Subjective, Objective, Assessment, and Plan) [30]. The resulting database contains a broad range of information, including indications and following prescriptions for therapy, referrals, hospitalization, and laboratory results. The structured information, such as diagnosis codes, is stored in a tabular format and the unstructured information is stored as free-text. Only the unstructured free-text was included in the corpus.

- *Specialist letters [SP]*

  These are letters written by a medical specialist – for example, a cardiologist – and they are also procured from the IPCI database. The purpose of these letters is to report back to the GPs after referral and consult in the hospital, updating them in relation to diagnostic deliberations and therapeutic strategies. These letters are in the form of scanned copies or summaries entered by the GP. These letters are also not always grammatically well formed.

- *Radiology reports [RD]*

  This set consists of the reports taken from the radiology department of the Erasmus Medical Center, The Netherlands. These reports contain descriptions and conclusions derived from diagnostic imaging as requested by medical specialists (doctors) or general practitioners. These reports are intended for communication between doctors and radiologists. The text is mostly generated by using an automatic speech recognizer (ASR) and therefore usually has proper grammar and structure by prevailing conventions of the Radiology department. The radiologists have the option to update the text generated by the ASR manually, which increases the probability of typos.

- *Discharge letters [DL]*

  This set consists of patient discharge letters taken from the Erasmus Medical Center. They serve a purpose comparable to the specialist letters in updating the GPs on everything that has occurred during the admission period including all outcomes and remaining problems. These letters are well formed because of their intended external use (by and beyond GPs) and continuity of care.

To select text from the above-mentioned sets, we first created a list of Dutch medical terms taken from the Unified Medical Language System (UMLS) [31]. The UMLS contains medical terms in 21 different languages, including Dutch. However, UMLS has limited coverage of terms in the non-English languages. From over 150 source vocabularies in the UMLS, only four contain Dutch language terms. Only UMLS terms belonging to one of 35 UMLS semantic types, mainly representing diseases, symptoms, and drugs, were included in the list. The final term list contains 153,573 Dutch terms, including synonyms and lexical variants that were present in the UMLS. For each of the four sets, documents containing at least one UMLS term were randomly selected to be included in the corpus. We used case-insensitive exact string matching to find the UMLS terms in the documents. Table 1 summarizes the characteristics of the four document types.

Table 1: Statistics of the four document types in the EMC clinical corpus

Type	**No. of documents**	**No. of recognized UMLS terms**	**No. of words per document***
GP entries	2000	3626	23 (14-38)
Specialist letters	2000	2748	39 (16-113)
Radiology reports	1500	3684	66 (46-94)
Discharge letters	2000	2830	163 (95-201)

* Median (interquartile range)

Each of the recognized terms in the corpus was annotated for the three contextual properties: negation, temporality, and experiencer. The definitions of the properties are adopted from the ConText algorithm [13].

- *Negation*

  This property has two values, 'Negated' or 'Not negated'. A clinical condition or term is labeled as 'Negated' if there is evidence in the text suggesting that the condition does not occur or exist, e.g., 'There was no sign of sinus infection', otherwise it is 'Not negated'.

- *Temporality*

  The temporality property places a condition along a time line. There are three possible values for this property: 'Recent', 'Historical', and 'Hypothetical'. A condition is considered 'Recent' if it is maximally 2 weeks old. Conditions that developed more than 2 weeks ago are labeled as 'Historical'. A condition is labeled as 'Hypothetical' if it is not 'Recent' or 'Historical', e.g., 'patient should return if she develops fever' [13].

- *Experiencer*

  Clinical text may refer to subjects other than the actual patient. The experiencer property describes whether the patient experienced the condition or someone else. For simplicity, we have defined only two possible values for this property: 'Patient' or 'Other', where 'Other' refers to anyone but the actual patient, e.g., 'Mother is recently diagnosed with cancer'.

The corpus was annotated by two independent annotators. They were provided with a guideline explaining the process and each of the contextual properties in detail, with examples. An expert who was familiar with all four types of clinical text resolved the differences between the annotators. The annotations were limited to the conditions previously identified using our custom Dutch UMLS terms. In The Netherlands, retrospective re- search with anonymized patient data does not fall under the scope of the WMO (Wet medisch-wetenschappelijk onderzoek met mensen ("Medical research involving human subjects act")), and does not have to be approved by a medical ethics committee For the IPCI data, the access was approved by the IPCI governance board (Raad van Toezicht).

We split each of the four document sets in our corpus into a development set and an evaluation set (50% each). The development set was used to tune the algorithm and the trigger lists. To account for possible overfitting of the algorithm on the development set, the performance of the algorithm was assessed on the evaluation set, which was used only for the final testing.

The ConText algorithm

The ConText algorithm [13], an extension of NegEx [9], is based on regular expressions and lists of trigger terms to determine the values of three contextual properties of a clinical condition: negation, temporality, and experiencer. The algorithm searches a sentence for triggers before or after the pre-indexed clinical condition. The default value of a property ('Not-negated' for

negation, 'Recent' for temporality, 'Patient' for experiencer) is changed if the condition falls within the scope of the trigger term. The default scope of a pre-trigger is from the right of trigger term to the end of the sentence, whereas the default scope of a post-trigger begins left- wards from the trigger term to the beginning of the sentence. The default scopes are overruled if a termination trigger is found before the end of the scope. For each property value (other than the default), the ConText algorithm maintains four lists of triggers: pre-triggers, post-triggers, termination triggers, and pseudo-triggers. Pre-triggers precede the location of a clinical condition in the text, e.g., *no signs of viral infection*. In this example, *viral infection* is the clinical condition and *no signs of* is the pre-trigger. Post-triggers follow a clinical condition, e.g., *viral infection is ruled out*. In this example, *ruled out* is a post-trigger. In both of these examples, the condition *viral infection* will be negated because it falls within the scope of the pre- and post-triggers. Termination triggers limit the scope of a pre- or post-trigger. Finally, there are phrases that look like triggers but do not act as such, e.g., *no change*. These are added to a pseudo-trigger list. The input to the algorithm is a sentence with marked clinical conditions. First, default values are assigned to the con- textual properties of each clinical condition. The default values are then updated using the following algorithm:

- Find all trigger terms (pre, post, pseudo, termination) in the sentence
- For each of the trigger terms found (from left to right)
    - If the term is a pseudo term, skip to the next term
    - Otherwise:
        - Find the scope of the trigger term
        - Assign appropriate contextual property values to all marked clinical conditions within the scope.

Several implementations of the ConText algorithm are available online [32].

ContextD: ConText for Dutch

The ConText algorithm uses pre-defined English trigger terms to determine the value of the contextual properties. We first attempted a fully automated translation of these triggers into Dutch using Google Translate [33], but the results appeared not to be comprehensive enough. A native Dutch speaker, who was also familiar with clinical texts, then checked all automatically translated terms, and added all possible variations of a trigger term.

The ContextD algorithm expects a sentence with marked conditions as its input. We used the Dutch sentence splitter in the Apache OpenNLP library [34] to split the text into sentences. Using our custom UMLS Dutch term list and case-insensitive exact string matching, we marked all the UMLS terms in the sentences.

ContextD works like the original ConText algorithm in using the trigger lists to find the values of contextual properties. The Java implementation of ConText [32] with the translated triggers was

used as a starting point. Using the development set, we iteratively refined the Dutch trigger lists and made a number of other modifications as described below:

GP specific rules

The general practitioners often negate the existence of a clinical condition by putting a minus sign after the term, e.g., *fever-*. We added a couple of rules to catch such occurrences (and their variations) of negation in the GP text.

Combined triggers

The value of a contextual property sometimes cannot be identified by a pre-trigger or a post-trigger alone, such as *nooit (never)* and *is weg (is gone)*. A similar weakness is also reported by Chapman et al. [9] for triggers not and no. For example, in the sentence '*Hij heeft verder nooit medicijnen gebruikt die de tinnitus beinvloeden (he has also never used medications that affect tinnitus)*', the trigger *nooit* is negating the use of medication but not the condition *tinnitus*. Some of the triggers translated from the English cannot be directly applied to the Dutch text because of the different word ordering in both languages. Such triggers have to be split before they can be applied. There are situations where a combination of two triggers is essential. Since there is no notion of de- pendency or connection between different trigger types in the original ConText algorithm, we introduced a few rules that look for a combination of triggers to be present in order to identify the correct value of a contextual property. For example, in the sentence '*Nooit urineweginfecties doorgemaakt*', the triggers *nooit (pre-trigger)* and *doorgemaakt (post-trigger)* combined suggest a negation for the term *urineweginfecties*. The pre-trigger *nooit* alone did not increase performance and hence was removed from the trigger list during the development.

Scope of trigger terms

ContextD uses different scopes depending on the trigger term. The default right-scope starts from the right of the trigger term and ends at the end of the sentence. The default left-scope starts leftwards from the trigger term and ends at the beginning of the sentence. We experimented with different scopes for different types of clinical text, which resulted in modifying the default scope for GP entries to 6 words and for specialist letters to 10 words. The default scope is overridden if a termination trigger appears before the end of the scope. For GP entries, which are mostly grammatically unstructured, some punctuation, such as comma and semicolon, were added as termination triggers to limit the scope of triggers. For specialist letters, only colon and semicolon were added to the termination triggers.

Temporality module

The original ConText algorithm has very few triggers to identify whether a clinical condition is historical. We added a temporality module that implements several regular expressions to look for evidence for historical events on both sides of the clinical term. An adjusted left and right scope was also implemented in the module to avoid getting false positives.

Evaluation

We computed precision (true positives/[true positives + false positives]), recall (true positives/[true positives + false negatives]), and F-measure (the harmonic mean of precision and recall: 2 * precision * recall/[precision + recall]) for each of the three contextual properties.

For the negation property, terms that were assigned the value 'Negated' were taken as the positive class and terms that were marked 'Not Negated' as the negative class. Similarly, for the experiencer property, terms marked as 'Patient' were taken as positives, 'Others' as negatives. For the temporality property, which has three values, each value was considered as the positive class against the other two combined (e.g. 'Recent' vs. 'Historical' and 'Hypothetical'). A true positive was defined as a term that was correctly assigned to the positive class, a false positive as a term that was incorrectly assigned to the positive class, and a false negative as a term that was incorrectly assigned to the negative class.

We used Cohen's kappa [35] to calculate the agreement between both annotators for each of the three contextual properties. Because the UMLS terms were already marked in the sentences, the inter-annotator agreement was calculated for the labels only.

RESULTS

This section provides the annotation results of the EMC Dutch clinical corpus and the performance of the ContextD algorithm.

Table 2 shows the inter-annotator agreement for each report type in the corpus. According to the Altman classification [36], kappa is very good for 'Negated' and 'Recent' values (with the exception of 'Recent' on Radiology reports, which is good), moderate-to-good for 'Historical' values, and moderate for *hypothetical* values. The kappa for the *experiencer* property is very good except for in the radiology reports where a moderate agreement is observed. The lowest kappa score (moderate) of 0.46 is observed for the specialist letters for the value *hypothetical*. Because there was no hypothetical term in the discharge letters, no kappa was calculated for this property value.

Table 2: Inter-annotator agreement on contextual properties in the EMC clinical corpus

Document type	Negated	Recent	Historical	Hypothetical	Patient
GP entries	0.90	0.86	0.57	0.48	0.92
Specialist letters	0.90	0.93	0.62	0.46	0.98
Radiology reports	0.93	0.61	0.63	0.57	0.53
Discharge letters	0.94	0.95	0.56	n/a	0.98

Table 3 shows, for each report type, the distribution of the values of the three contextual properties. The distribution of the negated terms does not vary much between different report

types. Historical conditions occur more frequently in specialists' letters and discharge reports, 8% and 6% respectively, in comparison to GP entries (2%) and radiology reports (3%). This can be explained by the fact that specialist and discharge letters often include descriptions of the patients' past medical history. Hypothetical terms are absent in discharge letters and infrequent (1% to 2%) in the other report types. The value other for the experiencer property is also found infrequently (from 0.1% to 2%) in all report types.

Table 3: Distribution of the contextual property values in different types of clinical documents

Document type	**Total**	**Negation**		**Temporality**			**Experiencer**	
		Negated	**Not-Negated**	**Recent**	**Historical**	**Hypothetical**	**Patient**	**Other**
GP entries	3626	12%	88%	97%	2%	1%	98%	2%
Specialist letters	2748	15%	85%	90%	8%	2%	99%	1%
Radiology reports	3684	16%	84%	96%	3%	1%	99.9%	0.1%
Discharge letters	2830	13%	87%	94%	6%	0%	98%	2%

Table 4 shows the number of English triggers used by the original ConText algorithm and the number of Dutch language triggers used by ContextD. About 60% of the English triggers were translated one-to-one into Dutch. For the remaining English triggers, several possible Dutch translations were added resulting in a much larger number of Dutch triggers. For example, for the English negation trigger 'never had', three equivalent Dutch triggers were added: 'nooit gehad', 'had nooit', and 'hadden nooit'. All the original triggers from ConText were translated to Dutch without particular issues.

Table 4: Number of English and Dutch trigger terms for each contextual property

Contextual property	**English triggers**	**Dutch triggers**
Negation	160	395
Temporality	42	62
Experiencer	44	52
Total	246	509

Table 5 shows the performance on the evaluation set of the ConText algorithm using only automatically and manually translated Dutch triggers (baseline) and of the ContextD algorithm after all modifications (final).

Table 5: Results on the evaluation set using only the translated terms from English to Dutch (baseline) and the final ContextD results with modifications (final)

		Precision		Recall		F-score	
Property value	**Total**	**Baseline**	**Final**	**Baseline**	**Final**	**Baseline**	**Final**
Negated							
GP entries	175	0.96	0.88	0.66	0.90	0.78	0.89
Specialist letters	177	0.93	0.84	0.63	0.90	0.75	0.87
Radiology reports	287	0.96	0.91	0.55	0.97	0.70	0.93
Discharge letters	180	0.98	0.92	0.67	0.93	0.79	0.92
Recent							
GP entries	1365	0.97	0.98	0.98	0.94	0.98	0.96
Specialist letters	919	0.91	0.95	0.99	0.92	0.95	0.94
Radiology reports	1341	0.97	0.98	0.98	0.96	0.97	0.97
Discharge letters	1140	0.93	0.97	0.98	0.91	0.95	0.94
Historical							
GP entries	28	0.15	0.17	0.17	0.54	0.16	0.26
Specialist letters	66	0.47	0.41	0.10	0.76	0.17	0.54
Radiology reports	52	0.30	0.37	0.30	0.67	0.30	0.48
Discharge letters	90	0.36	0.39	0.13	0.78	0.19	0.52
Hypothetical							
GP entries	17	0	0	0	0	0	0
Specialist letters	29	0	0.67	0	0.07	0	0.13
Radiology reports	6	0	0.67	0	0.33	0	0.44
Discharge letters	0	0	0	0	0	0	0
Patient							
GP entries	1379	0.98	0.98	1.00	0.99	0.99	0.99
Specialist letters	999	0.99	0.99	1.00	0.99	0.99	0.99
Radiology reports	1398	0.99	1.00	1.00	1.00	1.00	1.00
Discharge letters	1220	0.98	0.99	1.00	1.00	0.99	0.99

The baseline performance of the algorithm was poor on the historical terms and could not identify a single hypothetical term. The experiencer property was the easiest to assign, which is reflected in the high baseline performance. For the negation property, the precision was high for all report types but the algorithm missed many negated terms, i.e., recall was low. For the final ContextD, the recall was considerably improved for negation and historical values on all report types. Although the performance was improved for hypothetical values on specialist letters and radiology reports, overall it remained poor.

The ContextD algorithm utilized 23 unique triggers to identify *negated* terms, 5 unique triggers to identify *historical* terms, 3 unique triggers to identify *hypothetical* terms, and 10 unique triggers to identify other terms across all report types. Among the 23 unique triggers for the negation property, the trigger term 'geen' (no) was used most frequently. The most used triggers for the temporality property and the experiencer property were 'status na' (status after) and 'moeder' (mother), respectively.

Table 6 shows an analysis of 25 randomly selected false negatives for different contextual property values in the evaluation set. In 40% of the errors, the evidence trigger was missing from our trigger list. For instance, in the entry 'Fam.anamn blanco voor trombose…' (No family history for thrombosis…) the trigger *blanco voor* was missing, resulting in misclassifying the negated concept trombose as Not Negated. These errors can be prevented by adding triggers to the ContextD trigger lists. It is important to note here that some of the trigger terms causing these errors (e.g., *is weg [is gone]*) were intentionally not added in the triggers list to avoid too many false positives. In 19% of the errors, a pre- or post-trigger alone could not correctly identify the property value of a term. These errors may be prevented by rules that combine pre- and post-triggers along with the distance to the actual term (see Combined Triggers) or rules restricting the scope of a particular trigger. For example, in the sentence 'Ochtendstijfheid: nee Nachtelijk rugpijn: nee, Wel zonne-allergie…' (Morning stiffness: no nightly back pain: no, sun allergy present…), the concept *Ochtendstijfheid* could have been identified by adding: nee as a post-trigger with a maximum scope of 2 words. In 17% of the errors, the sentences were too complex to identify and generalize any trigger or pattern. For example, in the sentence 'flinke ruizen drukpijn colon erge pijn in flank sinds een aantal dagen dacht zelf aan niersteen advies' (significant wheezing pressure pain colon severe pain in flank since a few days was thinking of kidney stone advise), the concept *niersteen (kidney stone)* is hypothetical. The possible trigger *dacht zelf (thought himself or herself)* could not be used because of its negative impact in terms of false positives. In 8% of the errors, a variation of the trigger (e.g., a different verb form) was used. The remaining 16% of the errors were due to miscellaneous reasons, such as typos (e.g., no space between the trigger and other words), sentence splitting errors, or the trigger being in another sentence than the condition.

Table 6: Error analysis of false negatives in the evaluation set

Error	**Negated**	**Historical**	**Hypothetical**	**Patient**	**Total**
Missing trigger	15	7	7	11	40
Complex trigger	1	8	2	8	19
Complex sentence	1	-	15	1	17
Trigger variation	-	7	-	1	8
Other	9	3	1	4	16
Total	25	25	25	25	100

Table 7 shows an analysis of 25 randomly selected false positives for the different property values. The hypothetical and patient values had less than 25 false positives, so all those available

were included in the analysis. In 37% of the errors, the scope of the evidence trigger wrongly included the condition. For example, in the sentence 'Conclusie Geen oogheelkundige verklaring voor de hoofdpijn' (Conclusion No ophthalmologic explanation for the headache), the pre-trigger *Geen* is wrongly negating the concept *hoofdpijn* although it has a limited scope. Annotation errors caused 14% of the errors. Half these annotation errors were because the annotators failed to pick the historical trigger 'status na' (status after) resulting in those terms being labeled as either Recent or Hypothetical. Two ambiguous triggers for the experiencer property ('pa', which could mean 'dad' or 'pathology', and 'oma', which could mean 'grandmother' or 'acute otitis media') caused 14% false positives. Some of the regular expressions in our temporality module caused 11% of the errors because they were either not specific enough or were missing some variations in the text. For example, in the sentence *'… geen dyspnoe wel net influenza gehad ferro en vit c als <3 weken niet beter revisie…' (…no dyspnea recently had influenza ferro and vit c if <3 weeks not better revision…)*, the temporality module identified *3 weken (3 weeks)* close to the concept *influenza* and wrongly labeled it as historical. These types of errors could be avoided by looking for extra evidence such as *net (recently)* and relational operators such as < in combination with the time. In 9% of the false positives, the error was due to missing pseudo triggers. For example, in the sentence *'met requip niet minder krampen en wel zwabberig,…' (with requip no fewer cramps and also unstable,…)*, the pseudo-trigger *niet minder* was missing in the trigger list, resulting in wrongly classifying *krampen* as Negated. The remaining 15% errors were due to several other reasons.

Table 7: Error analysis of false positives in the evaluation set

Error	**Negated**	**Historical**	**Hypothetical**	**Patient**	**Total (%)**
Trigger does not apply to condition	9	7	8	8	32 (37)
Annotation error	2	8	2	-	12 (14)
Ambiguous trigger	-	-	-	12	12 (14)
Trigger problem	-	10	-	-	10 (11)
Missing pseudo trigger	8	-	-	-	8 (9)
Other	6	-	3	4	13 (15)
Total	25	25	13	24	87 (100)

Table 8 shows a comparison of the performance of the final ContextD algorithm and the original ConText algorithm. The original ConText algorithm was evaluated on six different English clinical document types [13]. For the comparison, we have selected two document types that appear similar in both studies. An absent precision or recall means that the results could not be calculated because the sum of true positives and false positives or the sum of true positives and false negatives was zero [13]. For the negation property, both algorithms have the same F-score for the radiology reports, but ContextD appears to perform somewhat better on the discharge

letters. For the historical property, no comparison could be made for the radiology reports since no F-score was provided for the ConText algorithm. For discharge letters, the ConText algorithm performs better. The low performance of ContextD is due to the high number of false positives (low precision) of which many are annotation errors. For the hypothetical property, no comparison on the same document type could be made since for the radiology reports no results were provided for the ConText algorithm, and for the discharge letters no hypothetical terms were present in the Dutch corpus. For the experiencer property, both algorithms performed equally well.

Table 8: Comparison of the original ConText algorithm for English with the adapted ContextD algorithm for Dutch. For ConText, the results are taken from [13]

		ConText (English)			**ContextD (Dutch)**		
Category	**Document type**	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Negation	Radiology reports	1.00	0.86	0.93	0.91	0.97	0.93
	Discharge letters	0.84	0.89	0.86	0.92	0.93	0.92
Historical	Radiology reports	-	-	-	0.37	0.67	0.48
	Discharge letters	0.68	0.77	0.73	0.39	0.78	0.52
Hypothetical	Radiology reports	-	-	-	0.67	0.33	0.44
	Discharge letters	1.00	0.92	0.96	-	-	-
Experiencer	Radiology reports	-	-	-	1.00	1.00	1.00
	Discharge letters	1.00	1.00	1.00	0.99	1.00	0.99

DISCUSSION

In this paper, we describe and evaluate ContextD, an algorithm to identify contextual properties of medical terms in Dutch clinical text. To develop and test ContextD, we have also created the EMC Dutch clinical corpus, with annotations for the three contextual properties negation, temporality, and experiencer.

The EMC Dutch clinical corpus covers four different types of electronically stored clinical text: entries from the general practitioner, radiology reports, and two sets of medical letters after outpatient treatment (i.e. specialists' letters) or hospital admission (i.e. discharge letters). The combination of these texts can be considered a representative selection of the documented medical process in the broadest sense, including the patient's first interactions with the general practitioner, referrals and advanced (imaging) diagnostics in the hospital, and ultimately reporting-back to the general practitioner after polyclinic consult or discharge after hospital admission.

Although the GP entries have the smallest size among the four document types in our corpus, they contain more UMLS terms than the discharge letters, which are the largest. This can be explained by the fact that our Dutch term list was small, containing mainly common clinical terms, which are more likely to be mentioned in GP records. The statistics shown in Table 1, therefore, do not give a realistic view on the occurrence and coverage of clinically relevant terms in different Dutch clinical texts. A more complete Dutch term list would have identified many more terms in the clinical text.

The corpus was annotated by two independent annotators. Looking at the differences between the annotators a few observations can be made. Medically schooled annotators are prone to using information outside the context and make considerations based on prior knowledge concerning the natural course of a condition. On various occasions, one annotator labeled a term as historical based on the assumed chronicity of the disease. At times, annotators had different opinions about keywords such as 'status na' (status after), which suggests a longer existing condition. One annotator considered such cases as a part of medical history and often labeled the terms as historical whereas the other annotator sometimes labeled the terms as recent and sometimes as historical because of the uncertain time frame. The annotators often differed on the assignment of hypothetical values to terms, e.g., for terms that were part of a differential diagnosis. In the sentence 'differentiaal diagnostisch werd gedacht aan appendicitis of diverticulitis' (for the differential diagnosis appendicitis and diverticulitis were considered), one annotator labeled appendicitis and diverticulitis as recent, reasoning that if they exist they exist now, whereas the second annotator labeled both terms as hypothetical. The inter-annotator agreement for 'Patient' is low for the radiology reports (cf. Table 2), which can be explained by the very low number of non-patients (class 'Other', see Table 3). With such highly imbalanced class distributions, even a small number of annotation disagreements can result in a low kappa value.

ContextD baseline results showed poor performance for 'Historical' and 'Hypothetical' values (cf. Table 5). The *recent* and *patient* values, which were the default values for the temporality and experiencer properties, showed good results. The final ContextD results (cf. Table 5) show the improvements especially for the negation and historical values. The most difficult category turns out to be the hypothetical value for the GP entries where the algorithm failed to correctly identify a single hypothetical value. Only few hypothetical terms were contained in the corpus, even less in the training set that we used to expand our trigger lists. We did not find many consistent patterns in the training set to identify hypothetical terms effectively. About a third of the errors in the evaluation set were due to the missing trigger 'bij' (upon), which did not occur in the training set. The rest of the errors were due to the sentences being too complex to identify and generalize a trigger or a pattern.

Although we had a much larger list of Dutch triggers compared to the English triggers, only a small number of trigger phrases accounted for the majority of the detected terms. This finding is consistent with findings in other languages [9, 26, 37]. Out of 395 possible Dutch triggers for the negation property, only 23 negation triggers were actually found in the evaluation set. The error analysis on the evaluation set suggested a number of new triggers to identify negations, historical, hypothetical, and experiencer property across all report types. Some of these triggers

were intentionally not included in the trigger lists because they decreased rather than improved performance on the development set. A similar problem of some triggers negatively affecting the result was also found in the Swedish study [26].

Although some automatic and linguistically motivated approaches exist to detect the scope [17,18,38], the default scopes used in ContextD are approximate due to lack of full grammaticality in the clinical text. Apart from the standard termination triggers, some additional constraints such as punctuations were added to limit the scope of triggers in GP entries and in specialist letters. The scope for negation was varied in length but never extended past the sentence boundary. Thus, negations that stretched over sentence boundaries were missed. The value of contextual properties may depend on the section of the clinical text, e.g., a symptom described in the previous history section will become historical regardless of how it is phrased. This information was not provided to the ContextD algorithm and as a consequence, terms may have been wrongly classified. As mentioned above, annotators sometimes used prior medical knowledge concerning the natural course of a condition to label a value of the contextual property, e.g., assigning historical value to a term for which the chronicity is assumed. Finding the right value for such terms is difficult for algorithms like ContextD, which rely solely on the information present in the direct neighborhood of the term. No effort was made in ContextD to separate patient-reported symptoms (complaints) and suspected diagnoses from the actual diagnoses made by the physician. The suspected diagnoses are usually hypothetical whereas symptoms and actual diagnoses are not, a distinction that requires understanding of the text and therefore is difficult to make for ContextD-like approaches. It is also important to note that the ConText algorithm is a simple algorithm meant to identify simple expressions using trigger lists, and was never expected to capture all attributes. We used case-insensitive exact string matching to find the UMLS terms in the documents. Any variation of a term such as a spelling mistake is likely to be missed by this approach. The same can also be true for the trigger terms. It is also to note that the terms with linguistic variability may occur in variable contexts, which may require some adjustments in the trigger scope or in the regular expressions.

The ContextD algorithm showed good performance in identifying negation and experiencer contextual properties. The performance for the historical and hypothetical (and even for negation and experiencer) properties can be further improved by adding new triggers found in the evaluation set. We observed some errors due to sentence splitting with Apache OpenNLP [34], which is trained on regular natural language text. Retraining the sentence splitter to work better with the Dutch clinical text, especially for the GP entries and specialist letters, would resolve some of the issues related to the missing context. The radiology reports and discharge letters are grammatically well structured; therefore, deep sentence parsing and using rule-based or machine-learning techniques to estimate the trigger scopes for these reports can be employed. To determine historical and hypothetical concepts better, it is important to incorporate information about the specific parts of clinical text (e.g., pre-history and diagnosis) in the algorithm. An extended assertion model that supports multiple values of negation is required to deal with speculation, e.g., the disagreements on diseases in the differential diagnosis.

CONCLUSIONS

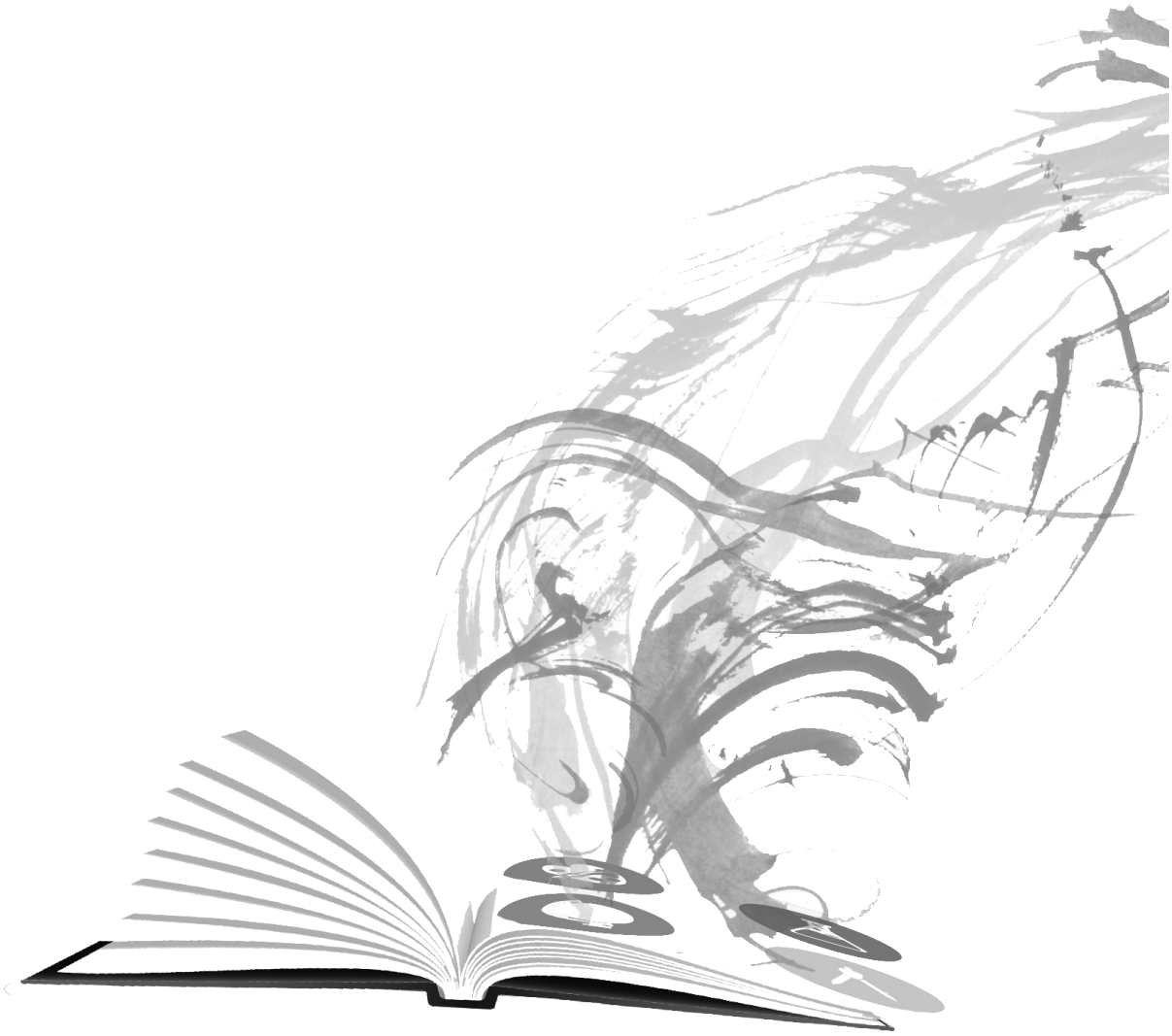
The ContextD algorithm showed good performance in identifying terms with negations and identifying who has experienced a particular medical condition across all four report types. The temporality property appears to be the most difficult one and methods to identify this property need to be further developed. The anonymized EMC Dutch clinical corpus, which was annotated for the three contextual properties negation, temporality, and experiencer, is the first publically available Dutch clinical corpus and can serve as a useful resource for further algorithm development.

REFERENCES

1. Jensen PB, Jensen LJ, Brunak S: **Mining electronic health records: Towards better research applications and clinical care.** *Nat Rev Genet* 2012, **13**(June)**:**395–405.
2. Friedman C, Hripcsak G: **Natural language processing and its future in medicine.** *Acad Med* 1999, **74**(8):890-895.
3. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB: **A general natural-language text processor for clinical radiology.** *J Am Med Informat Assoc* 1994, **1:**161–174.
4. Christensen LM, Haug PJ, Fiszman M: **MPLUS: A probabilistic medical language understanding system**. In *Proc ACL-02 Work Nat Lang Process Biomed domain -*, Volume 3. Morristown, NJ, USA: Association for Computational Linguistics; 2002:29–36.
5. Hahn U, Romacker M, Schulz S: **MEDSYNDIKATE – A natural language system for the extraction of medical information from findings reports.** *Int J Med Inform* 2002, **67:**63–74.
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: **Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications.** *J Am Med Inform Assoc* 2010, **17:**507–513.
7. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC: **MedEx: A medication information extraction system for clinical narratives.** *J Am Med Inform Assoc* 2010, **17:**19–24.
8. Zeng QT, 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 Inform Decis Mak* 2006, **6:**30.
9. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: **A simple algorithm for identifying negated findings and diseases in discharge summaries.** *J Biomed Inform* 2001, **34:**301–310.
10. Mutalik PG, Deshpande A, Nadkarni PM: **Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS.** *J Am Med Inform Assoc* 2001, **8:**598–609.
11. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, Wahner-Roedler DL: **A controlled trial of automated classification of negation from clinical notes.** *BMC Med Inform Decis Mak* 2005, **5:**13.
12. Huang Y, Lowe HJ: **A novel hybrid approach to automated negation detection in clinical radiology reports.** *J Am Med Inform Assoc* 2007, **14:**304–311.
13. Harkema H, Dowling JN, Thornblade T, Chapman WW: **ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports.** *J Biomed Inform* 2009, **42:**839–851.
14. Kilicoglu H, Bergler S: **Recognizing speculative language in biomedical research articles: A linguistically motivated perspective.** *BMC Bioinformatics* 2008, **9**(Suppl 11)**:**S10.
15. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, Montella D, Speroff T, Brown SH: **Detecting temporal expressions in medical narratives.** *Int J Med Inform* 2013, **82:**118–127.
16. Goldin I, Chapman W: **Learning to detect negation with "not" in medical texts**. In *Proc ACM-SIGIR 2003.*

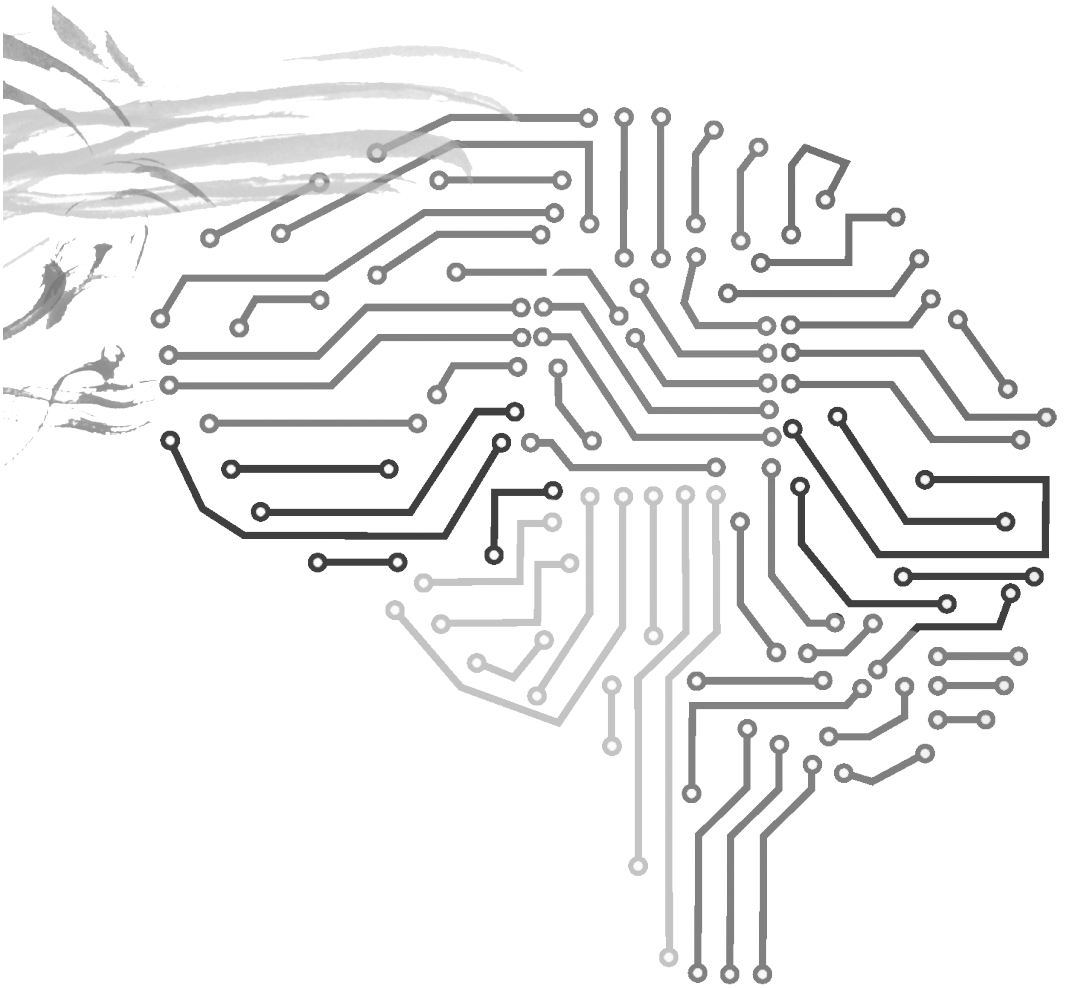
17. Agarwal S, Yu H: **Biomedical negation scope detection with conditional random fields.** *J Am Med Inform Assoc* 2010, **17:**696–701.
18. Morante R, Daelemans W: **A metalearning approach to processing the scope of negation.** *Proc Thirteen Conf Comput Nat Lang Learn - CoNLL'09* 2009, 21-29.
19. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J: **The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes.** *BMC Bioinformatics* 2008, **9**(Suppl 11)**:**S9.
20. Cruz Díaz NP, Maña López MJ, Vázquez JM, Álvarez VP: **A machine-learning approach to negation and speculation detection in clinical texts.** *J Am Soc Inf Sci Technol* 2012, **63:**1398–1410.
21. Light M, Qiu XY, Srinivasan P: **The language of bioscience: facts, speculations, and statements in between**. In *HLTNAACL 2004 Work BioLINK 2004 Link Biol Lit Ontol Databases*, Association for Computational Linguistics. Edited by Hirschman L, Pustejovsky J: 2004:17–24 [BIOLINK 2004 (Series editor)].
22. Velldal E: **Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature.** *J Biomed Semant* 2011, **2**(Suppl 5)**:**S7.
23. Goryachev S, Sordo M, Zeng Q, Ngo L: *Implementation and Evaluation of Four Different Methods of Negation Detection.* Technical report, DSG; 2006.
24. Uzuner O, Zhang X, Sibanda T: **Machine learning and rule-based approaches to assertion classification.** *J Am Med Inform Assoc* 2011, **16:**109–115.
25. Sun W, Rumshisky A, Uzuner O: **Evaluating temporal relations in clinical text: 2012 i2b2 Challenge.** *J Am Med Inform Assoc* 2013, **20**(Suppl 5)806-813.
26. Skeppstedt M: **Negation detection in Swedish clinical text: An adaption of NegEx to Swedish.** *J Biomed Semant* 2011, **2**(Suppl 3)**:**S3.
27. Deléger L, Grouin C: **Detecting negation of medical problems in French clinical notes**. In *SIGHIT Symp Int Heal informatics - IHI'12.* Edited by Proc ACM 2nd. New York, New York, USA: ACM Press; 2012:697–702.
28. ContextD: ConText for Dutch – Implementation and Resources, http://www.biosemantics.org/ContextD/
29. Vlug A, van der Lei J, Mosseveld B, van Wijk M, van der Linden P, MC S, van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project.** *Methods Inf Med* 1999, **38:**339–344.
30. Adequate Dossiervorming Met Het Elektronisch Patiëntendossier (ADEPD): *Nederlands Huisartsen Genootschap.* 2013:20–21.
31. Bodenreider O: **The unified medical language system (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32**(Suppl 1)**:**D267–D270.
32. **ConText implementation**. In: http://code.google.com/p/negex/downloads/.
33. **Google Translate**. In: http://translate.google.com.
34. **Apache OpenNLP library**. In: http://opennlp.apache.org/.
35. Cohen J: **A coefficient of agreement for nominal scales.** *Educ Psychol Meas* 1960, **20:**37–46 [Education and psychological measurement (Series editor)].
36. Altman DG: **Some common problems in medical research**. In *Pract Stat Med Res*: Volume Volume 1. London: Chapman & Hall; 1991:396–403.

37. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, Conway M, Tharp M, Mowery DL, Deleger L: **Extending the NegEx lexicon for multiple manguages**. In *MEDINFO 2013 - Proc 14th World Congr Med Heal Informatics.* Copenhagen, Denmark: Ios Press; 2013:677–681.
38. Apostolova E, Tomuro N, Demner-fushman D: **Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes**. In *Proc 49th Annu Meet Assoc Comput Linguist Hum Lang Technol short Pap - Vol 2.* Portland, Oregon, USA: Association for Computational Linguistics; 2011:283–287.



Chapter 3

Reducing feature dimensionality by normalizing text in electronic health records



ABSTRACT

Background

Clinical narratives that are found in the electronic medical records typically do not conform to any standard format. The use of standard and non-standard abbreviations, typographical errors, ill-formed and incomplete sentences makes automated methods to extract information challenging. This leads to large number of features for machine-learning tasks and further complicates the detection of clinical terms in narratives that may be relevant for pharmacoepidemiological purposes. This study aimed at methods to normalize text in electronic medical records as a way to reduce feature dimensionality.

Methods

We used IPCI (Integrated Primary Care Information), a Dutch general practitioners database, containing nearly 340 million free-text clinical narratives of more than 1.8 million patients. We used three methods to normalize text. In the first method, we normalized text by using standard terminologies such as MedDRA, MeSH, and SNOMED-CT. The second method focused on mapping abbreviations to their long-forms using a modified version of the Schwartz algorithm, and in the third method, we grouped words together based on their textual similarities and word lexemes. We used text normalization as a way to reduce feature dimensionality in a bag-of-words machine learning method and tested on two different clinical data sets.

Results

The IPCI database contained almost 6 million textually unique words, which followed a Zipfian frequency distribution. The coverage of the three terminology was low in the IPCI data as we were able to normalize only 272,791 (4.2%) words. We extracted 14,938 potential abbreviations using automated methods and mapped to their most probable long-forms from 148,281 candidate long-forms. Grouping similar words together reduced total unique number of words from almost 6 million to about 1.7 million. The normalized text resulted in 38% reduction in the features in one data set and 35% reduction in another data set. The reduction in features also showed positive impact on the classification performance.

Conclusions

We showed that the IPCI data has many term variations including spelling mistakes. We managed to reduce feature dimensionality using a word grouping based normalization approach. For IPCI like databases containing large number of clinical narratives, such normalization approach can be very useful as it can reduce the feature dimensionality for machine-learning tasks but also help other automated methods in better interpretation of the clinical text.

INTRODUCTION

General practitioners and specialists are increasingly using electronic medical records, rather than paper for keeping notes and information on the patient. Much information in these records is in clinical narratives (natural language). One problem in automated analysis of natural language is that there is no standard way of writing clinical narratives that are found in the electronic medical records (EMRs). The style of writing, the amount of information, and the use of language may vary from one healthcare center to another, even from one clinician to another and from one medical record information system to the other. Some clinical documents are used as a formal way of communication between clinicians such as discharge summaries or referrals; others may be used as references such as nurses' daily notes. In the Netherlands, the general practitioners (GPs) serve as a first point of contact to the patients. For every patient visit, the GP writes down the reason for visit, problems as described by the patient, an evaluation, and the resulting advice or prescription [1]. This information is mainly collected for primary care purposes and mostly serves as a diary to the GP. Since this information is for internal use only and the GPs are typically under time-pressure, free-text narratives in the primary care databases often contain ill-formed and incomplete sentences. EMRs may contain many typographical errors and standard and non-standard abbreviations. For brevity, we will refer to abbreviations and acronyms as short forms in the manuscript. Using such natural text is challenging since they complicate the detection of clinical terms such as drugs and their adverse effects, which may be relevant for pharmacoepidemiological purposes. For machine learning tasks where individual words (or combination of) are typically used as input variables or features, such noisy text results in a large number of features and due to this, feature selection models like bag-of-words suffer from the 'curse of dimensionality'.

Much work has been done in correcting typographical errors and normalizing short forms to their long-forms in general but only a few have focused on free-text in EMRs. Patrick et al. [2] used a dictionary-based heuristic approach for spelling corrections in clinical notes. They used several sources to build their dictionary. An edit distance algorithm was used to generate suggestions and a trigram model was then used to rank the suggestions. Crowell et all. [3] used the open-source GNU Aspell program to generate suggestions for misspelled words. They used frequencies of the possible corrections to rank the suggestions instead of using the default ranking by the tool. Tolentino et al. [4] built a dictionary from the Unified Medical Language System (UMLS) and WordNet sources for spelling corrections in vaccine safety reports. Kenneth et al. [5] used Shannon's noisy channel model to detect and correct misspellings in clinical free-text records. They also used a named entity recognition system to prevent person names from being corrected. Rohit J Kate [6] used a novel method to automatically learn patterns of variations of clinical terms from known variations from the UMLS. In non-English EHRs, Sikloski et al. [7] used a statistical method to correct single spelling errors in Hungarian clinical records. Grigonype et al. [8] used a combination of edit distance and phonetic similarity algorithms to generate suggestions and then used an n-gram model to analyze the suggested corrections for Swedish EHRs.

A comparative study [9] evaluated MetaMap [10], the Medical Language Extraction and Encoding System (MedLEE) [11], and cTAKES [12] on handling short forms in discharge summaries concluded that correct identification of clinical abbreviations is still a challenging task. Task 2 of the shared CLEF (Conference and Labs of the Evaluation Forum) eHealth challenge [13] focused on normalizing short forms to aid patient understanding of clinical text. The task organizers reported that reasonably high accuracy can be achieved on normalizing short forms but resolving ambiguous short forms is still challenging.

To our knowledge, there is no study that focused on normalizing words (clinical and non-clinical) in Dutch primary care EHRs. The effect of normalization on machine learning tasks, such as automated identification of patients with a certain disease or a symptom, has not been considered before as well. The amount of textual variations caused by typographical errors, and the use of standard and non-standard abbreviations that are often ambiguous make normalization a very challenging task. Most of the spelling correction methods are dependent on dictionaries and building a comprehensive medical dictionary is very difficult [5]. A large number of misspellings or morphological variations greatly increase the number of features in machine-learning tasks when words are used as features. To reduce these problems, variations and misspellings of a word should be normalized to one single representation. In this study, we aimed at methods to normalize text in Dutch EHRs. The normalization process involved using standard terminologies, grouping similar words together without using a dictionary, and automatically finding and mapping short-forms to their full-forms in the database. The normalization was used as a way to reduce feature dimensionality in a bag-of-words model. Performance was evaluated on two different clinical data sets.

METHODS

Data used in this study were taken from the Integrated Primary Care Information (IPCI) [14], a Dutch general practitioners database. The database is a collection of longitudinal electronic medical records from general practitioners in the Netherlands. The medical records contain medical notes related to symptoms, physical examinations, assessments and diagnoses, clinical findings, prescriptions and indications for therapy, information about patient referrals, hospitalization, and laboratory results. One patient record in IPCI may consist of one or more entries, where each record pertains to a patient visit or a letter from a specialist. In total, the database contains nearly 340 million free-text entries of more than 1.8 million patients.

Normalization approaches

We used three approaches to normalize text in the EHRs. In the first approach, we explored the possibility of normalizing words in the EHRs to standard words from a set of terminologies. The second approach was aimed at normalizing short-forms to their long-forms in an automated fashion. In our third approach, we normalized words based on their similarity to each other. The similarity was measured using an edit distance algorithm and word lexemes.

Use of terminologies

We selected three commonly used terminologies, MedDRA (Medical Dictionary for Regulatory Activities), MeSH (Medical Subject Headings), and SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) for this approach. The Dutch translations of concepts and terms were extracted from the UMLS. The terminologies were supplemented by automatic translation of English terms using Google Translate for concepts without a standard Dutch translation in the UMLS. No context was available for the terms to use during translation; therefore, it is possible that some of the terms are not translated accurately. For terms where the translator returned multiple alternatives, we picked the first one.

We used Peregrine, our dictionary-based concept recognition system [15], to compute the coverage of the terminologies in the IPCI text. In short, Peregrine uses a user-supplied dictionary or terminology and splits the terms in the terminology into sequences of tokens. When such a sequence of tokens is found in the text (using exact matching), the term and the concept associated with that term is recognized.

Mapping short-forms to long-forms

Free-text entries in IPCI contain many short-forms, most of which are non-standard and ambiguous. They either are created by the GPs on the fly, known only to them, or have specific meaning in one clinical domain. Identifying short-forms and their long-forms (definitions) is a challenging task [9,13], especially from ungrammatical free-text. We first identified potential short-forms in text by automated filtering; from a list of all unique words in the text, we first selected words consisting of a maximum of four characters (excluding the punctuations). All the dictionary and stop-words were then removed from the selected words. We used the Dutch stop-word list available in Snowball [16], a library of stemming algorithms. The remaining words were considered as potential short-forms. A modified version of the Schwartz algorithm [17] was then applied on the entire IPCI database entries to identify long-form candidates for each of the potential short-forms. The Schwartz algorithm is a simple algorithm that was developed to detect short-form long-form pairs from biomedical text. The algorithm requires both short-form and the long-form in the same sentence. In GP notes, short-forms and long-forms are hardly found together in the same sentence. The original Schwartz algorithm was therefore modified to take in a list of short-forms and then look for long-forms in all text.

To reduce noise, all long-form candidates containing a digit or a special character were removed. For each short-form long-form pair, we identified a type: whether a short-form was a truncation of the long-form, an acronym or initialism (abbreviation consisting of initial letters pronounced separately), portmanteau (blend of two or more words), or something else. We removed all short-forms of length 2 because of the very high number of resulting noisy long-forms identified by the algorithm. If there were still multiple potential long-forms for an abbreviation after filtering, the long-form with the highest frequency in the database was chosen as the final long-form for the abbreviation.

Word groupings

We grouped words together based on their textual similarities in order to normalize all word variations into a single representation. Two different methods were used for the word groupings.

Word groupings based on edit distance

We first created a large similarity matrix of all unique words in IPCI by using the Damerau-Levenshtein distance algorithm, which counts for four operations: insertion, deletion, substitution, and transposition. The Damerau-Levenshtein distance algorithm differs from the widely used Levenshtein distance algorithm in that it allows for transposition. For each pair in the similarity matrix, the Damerau-Levenshtein distance between the two words and the frequency of each word in the IPCI database was recorded. To create the word groupings or clusters, we used an approach similar to the Partitioning Around Medoid (PAM) algorithm [ref], which is one of the approaches in *k*-medoid clustering. In PAM, *k* points are initially selected to be the cluster medoids. Each of the remaining points are then associated to the cluster with the closest medoid. In the last step, members of the clusters are swapped if that decreases the total cost of the clusters. We used a slightly different approach for clustering. We first created a list of all unigrams in IPCI and sorted them on their frequency. Instead of pre-selecting *k* words as medoids, we processed the sorted list from top to bottom one by one and created clusters around the medoids. The cluster-creating algorithm is described in pseudocode in Table 1.

Table 1: Pseudocode algorithm for creating clusters

```
Input: L, frequency-sorted list of words; S, matrix of similarities between words
Output: C, ...

X = empty set
For each word Wi in sorted list L
    If word Wi is not in the list X of already consumed words
    Cx = empty set
    Set word Wi as medoid of Cx
    Extract all precomputed values R from the similarity matrix for Wi
    For each word Rj in R
        If distance between Wi and Rj is less than or equal to the pre-specified criteria
            If additional_criteria also matched
                Add word Rj to cluster Cx
                Add word Rj to the list X
    Select Wi as the head word of the cluster Cx
```

We generated four sets of clusters based on different criteria, as shows in Table 2. The headword is used as the normalized word to represent all the words in the cluster.

Table 2: Member selection criteria for different clustering schemes

Clustering scheme	Criteria
Baseline	IF Length of medoid word <= 5 characters THEN Maximum allowed distance 1 character IF Length of medoid word > 5 AND <= 7 characters THEN Allowed distance 2 characters IF Length of medoid word > 7 AND <= 10 characters THEN Allowed distance 3 characters IF Length of medoid word >10 characters THEN Allowed distance 4 characters
Scheme 1	IF Length of medoid word < 5 characters THEN Allowed distance 0 character IF Length of medoid word >= 5 AND <= 7 characters THEN Allowed distance 1 character IF Length of medoid word > 7 AND <= 10 characters THEN Allowed distance 2 character IF Length of medoid word >10 characters THEN Allowed distance 3 character
Scheme 2	All criteria of Scheme 1 AND Frequency of the member word in the cluster <= 1000
Scheme 3	All criteria of Scheme 2 AND Frequency of the member word <= half of the frequency of the medoid word

Word groupings based on lemmas

The second method we used to create word clusters was based on word lemmas. We used Frog software [18] to find lemmas of all the words in the IPCI database. Frog is a memory-based morphosyntactic tagger and dependency parser for the Dutch language and it is freely available. Besides lemmatization, Frog can tokenize, tag, morphologically segment word tokens, identify named entities, and assign dependency graphs to Dutch text. To generate clusters, we grouped all words with the same lemma together. For each cluster, the lemma was used as the head word.

Evaluation

We performed two different types of evaluations in this study. The first evaluation focused on evaluating different clustering schemes in order to select the best one. Second to evaluate the effect of word normalization as a feature reduction method in a machine-learning task.

Clustering can be evaluated using an internal or an external validation method [19]. An external validation requires information that is not present in the data, usually a gold standard or a reference set. An internal validation only relies on the information present in the data. In the absence of a reference set to compare the quality of the clustering, we used an internal clustering

validation method. Several internal validation methods are described in [19–21]. We used Davies-Bouldin Index (DBI) [22] as an evaluation measure. DBI measures the separation between the clusters and the compactness within the clusters based on the distances. We measured separation by calculating the distance between the centroids of the clusters and compactness by measuring the distance of the cluster members to its centroid. Ideally, the separation between the clusters should be as large as possible and the within cluster scatter as low as possible.

We used two clinical data sets to evaluate the effect of word normalization on classification performance. The data for each set was taken from the IPCI database. Both data sets were manually validated for positive and negative cases. The process of extracting and validating clinical data is explained elsewhere [23]. The first data set 'colorectal cancer' (CRC) consisted of 4521 cases, out of which 1946 cases were positive and 2557 cases were negative. The second data set 'colorectal polyps' (POYLP) consisted of 4502 cases, out of which 926 were positive and 3576 were negative. We used C4.5, a well-known decision tree learner in our experiments.

RESULTS

The IPCI database contained almost 6 million unique words. Figure 1 shows the frequencies of all IPCI word lengths in character. The frequency distribution of the unique words followed Zipf's Law, according to which the frequency of a word is inversely proportional to its rank (Figure 2).

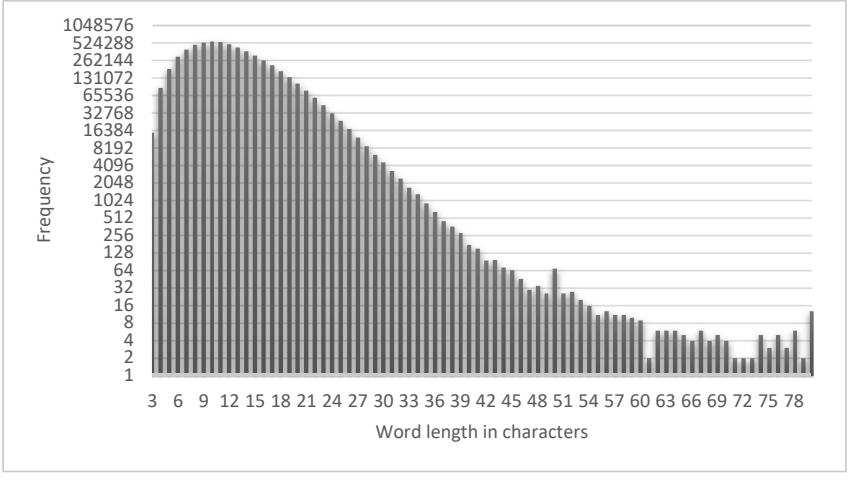


Figure 1: Histogram of lengths of words in the IPCI database

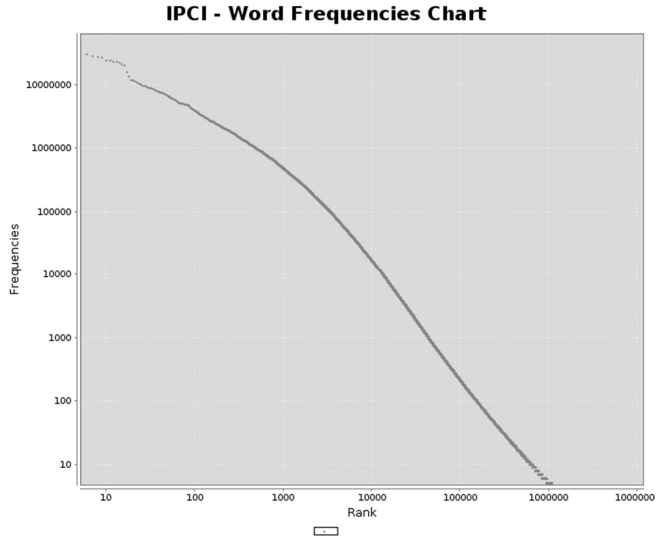


Figure 2: IPCI word frequency chart on a log-log scale. Horizontal axis represents word rankings according to frequency; vertical axis represents word frequencies in the IPCI database

Table 3: Characteristics of the three Dutch-translated terminologies

Terminology	Dutch terms extracted from UMLS	Translated	Total Terms
MedDRA	71,185	1,442,224	1,513,409
MeSH	39,729	2,291,832	2,331,561
SNOMED-CT	0	2,650,162	2,650,162

Table 3 summarizes the characteristics of the terminologies. We computed the coverage of the three terminologies with Dutch medical terms on the IPCI data. Table 4 shows the coverage results on term level for each terminology.

Table 4: Coverage of the three Dutch terminologies

	IPCI Terms Found (Total)			
	MedDRA	**MeSH**	**Snomed-CT**	**Total**
Terms	116,666 (1,513,409)	89,441 (2,331,561)	66,684 (2,650,162)	272,791 (6,495,132)
Coverage (%)	7.71	3.84	2.52	4.2

The terms were identified using our Peregrine indexing engine. The three terminologies combined contained 6,495,132 non-unique terms. Out of 5,973,858 unique terms in IPCI, only 272,791 terms were found in any of the three terminologies. The overall coverage of IPCI terms in the terminologies was only 4.2%, with the highest of 7.71% in MedDRA and the lowest of 2.52% in Snomed-CT.

We first identified 14,938 potential short-forms in the database by using the method explained above. Later, the Schwartz algorithm was used to identify long-forms for each of the potential short-forms. The algorithm was applied to the whole IPCI database containing nearly 340 million free-text entries. For 14,938 short-forms, the algorithm identified 148,281 potential long-forms.

Table 5: Potential short-forms, their identified long-forms, and filtering results

Short form Length	**Short forms**	**Abbr. Long Form**	**F1: Same Long Form as Abbreviation (%)**	**F2: Long Forms Containing Digit (%)**	**F3: Long Forms Containing Special Characters (%)**	**F4: Long Forms Containing Stop-words (%)**
3	4774	79676	2019 (2.5)	647 (0.8)	10120 (12.7)	2303 (2.9)
4	10164	68605	2874 (4.2)	465 (0.7)	9202 (13.4)	8096 (11.8)
Total	14938	148281	4893 (3.3)	1112 (0.8)	19322 (13)	10399 (7)

Several filters were applied to identify and remove the incorrect long-forms identified by the algorithm. Table 5 shows the results of the short-form identification, Schwartz algorithm, and several filters applied on the identified long-forms. Sometimes the long-form identified by the algorithm was the same as the short-form, e.g., a.d.o. and A.D.O.; such obviously erroneous pairs were removed from the list. The long-forms containing one of the following characters were also removed: "\ , ; : \ - > < ( ) $ % ^ # ' & ? +]".

All potential short-forms and their long-forms pairs where the short-forms were not acronyms, initialisms, or portmanteaus (i.e. blend of words) were also removed. Table 6 shows the final number of short-forms and their long-forms after filtering.

In total, short-form filtering resulted in about 80% reduction and long-form filtering resulted in about 93% reduction.

Table 6: Filtered short-forms and their long-forms

Short-form Length	# of Short-forms	Short-form Reduction (%)	Long Forms	Long Forms Reduction (%)
3	1067	67	4203	95
4	1877	81	5645	92
Total	2944	80	9848	93

A histogram showing the distribution of short-forms and their long-forms before and after filtering is presented in Figure 3.

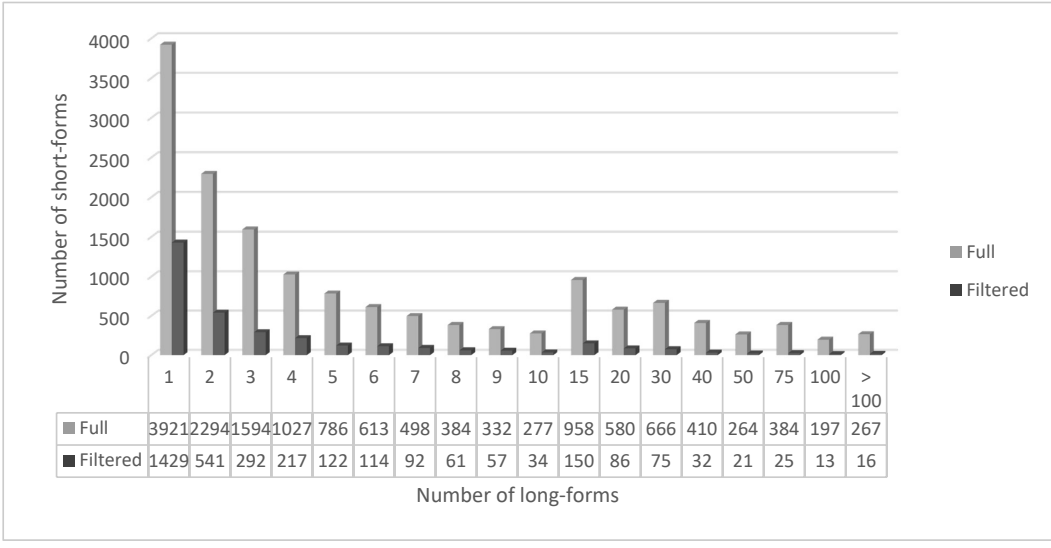


Figure 3: : Histogram showing number of short-forms and their long-forms for both full and filtered set

The Davies-Bouldin Index (DB Index) was used to evaluate and compare the different clustering schemes. (A lower DB Index indicates a better clustering scheme.) The baseline method for edit distance based clustering yielded 721,014 clusters with a DB Index of 1.54. Ignoring words smaller than 5 characters for similarity measures and increasing the error threshold (Scheme 1) immediately improved clustering performance but also increased the number of clusters to 1,553,557. Filtering clustering members on their frequencies (Scheme 2 and 3) further improved the clustering performance. The lowest DB Index score of 0.90 with 1,719,394 clusters was

observed for scheme 3. The second method, where clusters were generated using word lemmas, resulted in 5,099,050 clusters and a DB index of 5.18, much larger than the DB indices of the edit distance based methods.

Table 7: Number of resulting clusters for each clustering method and the corresponding DB Index score

Clustering method		**Number of Clusters**	**DB Index**
Edit Distance Based	Baseline	721014	1.54
	Scheme 1	1553557	0.96
	Scheme 2	1539575	0.95
	Scheme 3	1719394	0.90
Lemma Based		5099050	5.18

To see the effect of word normalization on feature reduction, we used the POLYP and CRC data sets. A bag-of-words (BoW) feature representation scheme was used on both sets. In total, 346539 features were extracted from the POLYP data set (Table 8). Several pre-processing filters were applied before extracting the features, such as removing numbers, words containing digits, and negated words. We used word clusters from scheme 3 of edit distance based clustering and lemma-based clusters for word normalization.

Table 8: Effect of word normalization on feature reduction using word clusters on two data sets

Data Set		**Edit Distance Based Clustering (Scheme 3)**		**Lemma Based Clustering**	
	# Features before normalization	# Features after normalization	Reduction	# Features after normalization	Reduction
POLYP	346539	214105	38%	329610	5%
CRC	292700	191893	35%	284715	3%

The normalization using edit distanced based clusters resulted in 38% reduction in the features whereas only 5% reduction was observed when lemma based word normalization was applied. A similar trend was observed on the second data set of CRC, where edit distance achieved 35% reduction as compared to only 3% reduction using lemma based clusters. The results from lemma normalization are not surprising considering the number of clusters it generated, due largely to the amount of textual variations such as typos, misspellings, and words without spaces in the text.

We used C4.5 classifiers on both data sets to see the impact of feature reduction on classification performance of identification of cases of CRC or polyps in the data set. The precision, sensitivity (recall), specificity, and F-scores were calculated for each of the data sets first without normalization and then normalization based on edit distance clusters and lemma-based clusters.

Table 9: Classification performance on both data sets with and without word normalization

Data Set		Precision	Sensitivity	Specificity	F-score
POLYP	Without Normalization	0.812	0.857	0.849	0.834
	With Edit Distance Based Normalization	0.808	0.879	0.841	0.842
	With Lemma Based Normalization	0.827	0.863	0.863	0.845
CRC	Without Normalization	0.880	0.829	0.971	0.854
	With Edit Distance Based Normalization	0.881	0.830	0.971	0.855
	With Lemma Based Normalization	0.883	0.825	0.972	0.853

The results in Table 9 show that feature reduction using word normalization (both methods) slightly improved the classification performance. On the POLYP data set, the sensitivity was increased from 0.857 to 0.879. However, the best F-score of 0.845 was achieved using lemma-based feature reduction. On the CRC data set, edit distance based feature reduction achieved the highest sensitivity of 0.830 and the highest F-score of 0.855. On both sets edit distance based feature reduction resulted in improved sensitivity.

DISCUSSION

In this study, we considered reducing feature dimensionality by normalizing text in the electronic health records and tested the impact of this on identification of certain cases from text. We showed that word clustering may be used to normalize text in medical record like databases which contains a lot of textual variations such as grammar, typographical errors, and words without spaces. We clustered textually similar words using Damerau-Levenshtein distance algorithm and using word lemmas. We applied feature reduction using both clustering methods on two data sets and observed improvement in the classification performance.

We identified almost 6 million textually unique words by processing nearly 340 million patient entries. We used three common terminologies containing Dutch translations from the UMLS to check their coverage in the IPCI database. We explored the possibility to use only identified medical terms as features in order to reduce feature dimensionality but we observed a very low term level coverage (4.2%). One reason could be that although the three terminologies we used are common but these were not perhaps not very suitable for this database. Many terminologies in the UMLS do not have Dutch translated terms and it is expensive to use machine translation on large scale to translate each English term into Dutch. Since the translation is also done without context, it is also possible that some of the terms are not correctly translated. It is hard to quantify whether using more terminologies will result in much different coverage scores. Another likely scenario for low coverage is the type of free text in the IPCI database. The text is often noisy as explained earlier and the Peregrine concept recognition system used in the experiments only does exact matching. Although Peregrine also uses stemming (i.e. reducing words to their stem or root form) in the matching process, it was not able to identify many words with large variations. Some of the typos such as spelling errors in the middle or at the beginning could have been captured using fuzzy matching, but it was not available in Peregrine.

There has been a lot of work done on identifying and normalizing short forms in biomedical text but not so much from the clinical text and even less from non-English languages. Clinical texts such as nurses notes and GP entries are written under time pressure and contain many standard and non-standard short-forms. Normalizing short-forms to their full form (long-form), which would improve feature extraction, is a challenging task in clinical text [9,13]. Most of the approaches involved supervised machine learning methods (such as SVM or CRF) where a manually labeled training set is usually available. In this study, we first used heuristics to identify potential short forms and then used the Schwartz algorithm to identify potential long-forms for each potential short-form. This simple method can essentially be used without worrying about the type of the text and the language. Since this is an unsupervised approach applied to a Dutch GP database, it cannot be directly compared with previous methods used. Implementing efficient post-processing filters is crucial since this method is prone to generate many false positives. We focused on the short-forms of lengths up-to 4 characters (excluding punctuations). This resulted in a large number of potential short-forms (cf. Table 5) and in a very large number of potential long-forms. Although various filters removed about 80% of the short-forms and about 93% of the long-forms (c.f. Tables Table 5 and Table 6), there were still many short-form/long-form pairs left. Many short-forms resulted in more than one long-forms (cf. Figure 3). Although others have considered this a word sense disambiguation (WSD) problem and tried to tackle it accordingly [9,13], we used a naïve approach to select the most frequent long-form among the options.

An edit distance and word lemma based methods were explored to cluster textually similar words together. The challenge with edit distance based methods is to define an optimal similarity threshold. We used a rather lenient threshold to calculate a baseline. Our results show that using a strict threshold improves clustering performance but at the cost of the resulting number of clusters. Based on our analysis of a small random set of clusters, we used word frequency information as an additional criterion to add words in the same clusters. For example, the edit distance between the Dutch words ‘specialisme’ (‘specialty’ in English) and ‘specialist’ (same in English) is two and using Scheme 1 they both ended up in the same cluster, which is semantically incorrect. We reasoned that the frequency of a certain typo or a misspelled word should not be very high and if a cluster member has a very high frequency in the database, it may actually be a correct word. Our results show that adding a criterion that the frequency of a cluster member must be lower than half of the frequency of the cluster head word, works best, achieving the best Davies-Bouldin Index of 0.90. Lemma-based clustering resulted in more than 5 million clusters from roughly 6 million words, indicating that the Frog lemmatizer was not able to assign the canonical forms to most words. This was not surprising for two reasons: 1) the Frog lemmatizer is trained on a non-clinical Dutch lexicon so it may not work as good on clinical terms, and 2) the text in the IPCI database is very noisy, and since lemmatization process usually involves vocabulary and morphological analysis of the words, it may have a strong impact on the performance. The edit distance based clustering method (scheme 3) reduced the total number of unique words by 71% as compared to the lemma based clustering method, which resulted in a total reduction of only 15%.

Previous studies on normalizing clinical data mainly focused on spelling corrections and abbreviation identification and expansion [3-10]. Most of those methods are dependent on

linguistic resources such as dictionaries or domain specific terminologies. None of the previous studies have looked into the effect of these normalizations on a machine-learning task. We considered text normalization as a way to reduce feature dimensions in medical record databases that contain plenty of word variants, and tested the impact of this on identification of diagnoses in this type of textual data. On both clinical data sets used in this study, POLYP and CRC, a large feature reduction was observed using the edit distance based feature normalization method as compared to the lemma based word normalization. The feature reduction had a small positive impact on the classification performance. On the POLYP data set, the highest F-score was observed for the classifier where features were normalized using word lemma based clusters. On both sets edit distance based feature reduction resulted in improved sensitivity, which is usually important for clinical data sets.

There were several study limitations. Firstly, the Schwartz algorithm identified many potential long-forms for each potential abbreviation. Although filtering removed many pairs, we still end up with many pairs. No efforts were made to disambiguate short forms and their long-forms. Rather, a simple approach was used to select the long-form using the frequency information. This may have resulted in picking up several incorrect long-forms. Secondly, for practical reasons only a couple of similarity thresholds were experimented with for edit distance algorithm. Thirdly, the frequency information used while clustering words may have resulted in incorrectly filtering some common and frequent grammar variations of words. For example, 'huisarts' (general practitioner) and the plural 'huisartsen' (general practitioners) ended up in two different word clusters. Since it would have required a lot of effort to train the lemmatizer on IPCI data, we choose to use Frog with its pre-trained models for lemmatization, which are not optimized for clinical text.

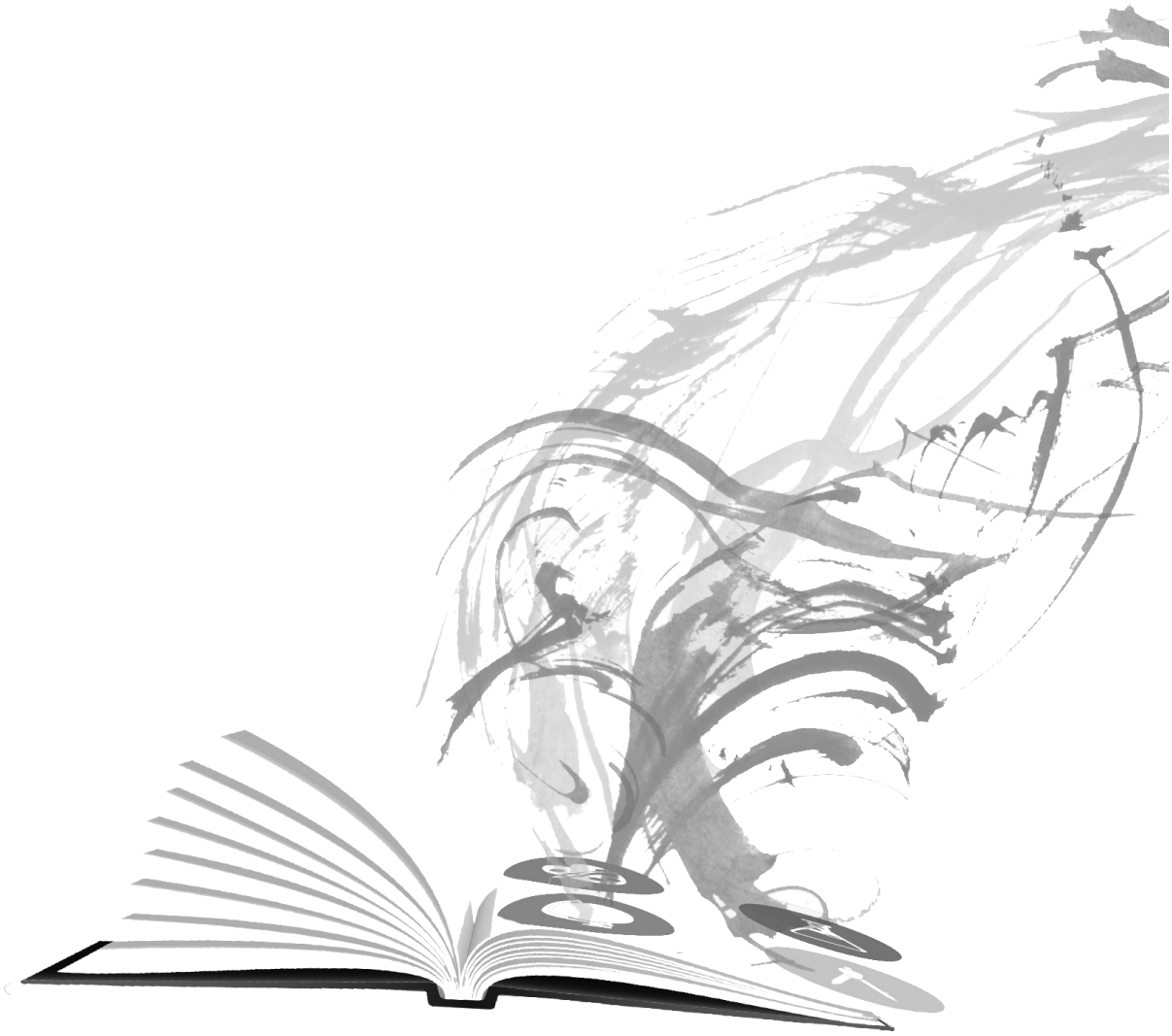
In conclusion, we managed to reduce feature dimensionality using a word clustering based normalization approach. We showed that word normalization resulted in better classification performance to identify diagnoses in this medical record dataset, especially in improving sensitivity. Finding several long-forms for each potential short-form indicates the ambiguity of short-forms used by GPs. Our results suggest that more efforts are required for better context-aware disambiguation. For IPCI like databases where the text is often grammar-free and contains a lot of variation, a feature normalization approach using textually similar word clusters could be very useful. Grouping similar words to get potentially more useful terms, identifying standard and non-standard abbreviations, and using standard terminologies are needed to better facilitate automated interpretation of the clinical text.

REFERENCES

1. *Adequate Dossiervorming Met Het Elektronisch Patiëntendossier (ADEPD)*. Nederlands Huisartsen Genootschap; 2013.
2. Patrick J, Sabbagh M, Jain S, Zheng H: **Spelling correction in clinical notes with emphasis on first suggestion accuracy**. In *2nd Work Build Eval Resour Biomed Text Min*; 2010(March):2–8.
3. Jonathan Crowell, MS, Qing Zeng, PhD, Long Ngo, PhD, and Eve-Marie Lacroix M: **A frequency-based technique to improve the spelling suggestion rank in medical queries**. *J Am Med Informatics Assoc* 2004, **11**:179–185.
4. Tolentino HD, Matters MD, Walop W, Law B, Tong W, Liu F, Fontelo P, Kohl K, Payne DC: **A UMLS-based spell checker for natural language processing in vaccine safety**. *BMC Med Inform Decis Mak* 2007, **7**:3.
5. Lai KH, Topaz M, Goss FR, Zhou L: **Automated misspelling detection and correction in clinical free-text records**. *J Biomed Inform* 2015, **55**:188–195.
6. Kate RJ: **Normalizing clinical terms using learned edit distance patterns.** *J Am Med Inform Assoc* 2015.
7. Siklosi B, Novak A, Proszeky G: **Context-aware correction of spelling errors in Hungarian medical documents**. In *Int Conf Stat Lang Speech Process*. Springer Berlin Heidelberg; 2013(July):248–259.
8. Grigonyte G, Kvist M, Velupillai S, Wirén M: **Improving readability of Swedish electronic health records through lexical simplification: First results**. In *Proc 3rd Work Predict Improv Text Readability Target Read Popul*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014:74–83.
9. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H: **A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries.** *AMIA . Annu Symp proceedings AMIA Symp* 2012, **2012**:997–1003.
10. Aronson AR, Lang F-M: **An overview of MetaMap: Historical perspective and recent advances**. *J Am Med Informatics Assoc* 2010, **17**:229–236.
11. Friedman C, Alderson P, Austin J, Cimino J, Johnson S: **A general natural-language text processor for clinical radiology**. *J Am Med Informatics Assoc* 1994, **1**:161–174.
12. Savova GK, Masanz JJ, Ogren P V, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: **Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications.** *J Am Med Inform Assoc* 2010, **17**:507–513.
13. Mowery DL, South BR, Christensen L, Leng J, Peltonen L-M, Salanterä S, Suominen H, Martinez D, Velupillai S, Elhadad N, Savova G, Pradhan S, Chapman WW: **Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth challenge 2013, Task 2**. *J Biomed Semantics* 2016, **7**:43.
14. Vlug A, van der Lei J, Mosseveld B, van Wijk M, van der Linden P, MC S, van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project**. *Methods*

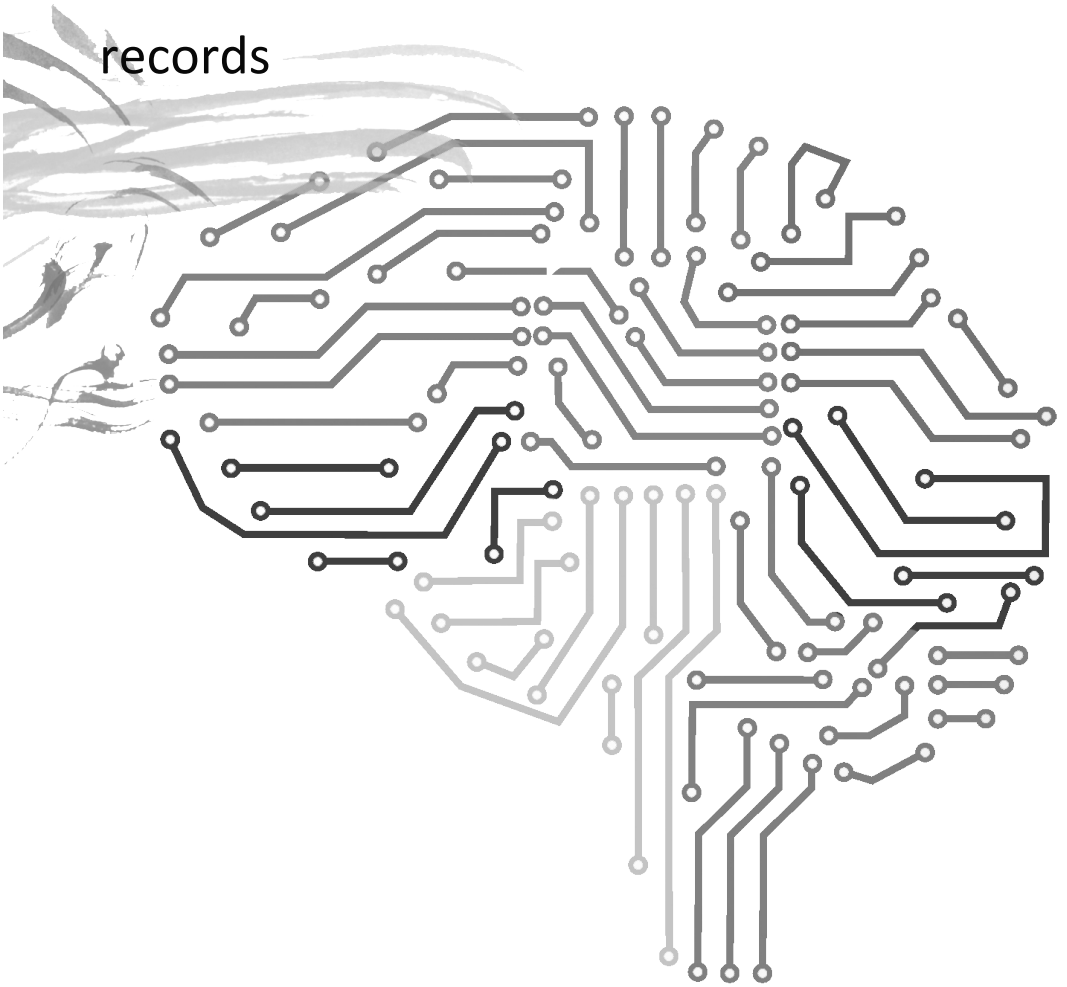
Inf Med 1999, **38**:339–344.

15. Schuemie MJ, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup**. In *Proc Second BioCreative Chall Eval Work*; 2007:131–133.
16. **Snowball stemmers**. :http://snowball.tartarus.org/.
17. Schwartz AS, Hearst MA: **A simple algorithm for identifying abbreviation definitions in biomedical text.** *Pac Symp Biocomput* 2003:451–462.
18. van den Bosch A, Busser B, Canisius S, Daelemans W: **An efficient memory-based morphosyntactic tagger and parser for Dutch**. In *Sel Pap 17th Comput Linguist Netherlands Meet*. Edited by Eynde F V, Dirix P, Schuurman I, Vandeghinste V. Leuven, Belgium; 2007:99–114.
19. Rendón E, Abundez I, Arizmendi A, Quiroz M E: **Internal versus External cluster validation indexes**. *Int J Comput Commun* 2011, **5**.
20. Liu Y, Li Z, Xiong H, Gao X, Wu J: **Understanding of internal clustering validation measures**. In *2010 IEEE Int Conf Data Min*. IEEE; 2010:911–916.
21. Jegatha Deborah L, Baskaran R, Kannan A: **A survey on internal validity measure for cluster validation**. *Int J Comput Sci Eng Surv* 2010, **1**:85–102.
22. Davies DL, Bouldin DW: **A cluster separation measure**. *IEEE Trans Pattern Anal Mach Intell* 1979, **1**:224–227.
23. Afzal Z, Schuemie MJ, Van Blijderveen JC, Sen EF, Sturkenboom MC, Kors JA: **Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records**. *BMC Med Inform Decis Mak* 2013, **13**.



Chapter 4

Generating and evaluating a propensity model using textual features from electronic medical records



ABSTRACT

Background

Propensity score (PS) methods are commonly used to control for confounding in comparative effectiveness studies. Electronic health records (EHRs) contain much unstructured data that could be used as proxies for potential confounding factors. The goal of this study was to assess whether the unstructured information can also be used to construct PS models that would allow to properly deal with confounding. We used an example of coxibs (Cox-2 inhibitors) vs. traditional NSAIDs and the risk of upper gastro-intestinal bleeding as example, since this association is often confounded due to channeling of coxibs to patients at higher risk of upper gastro-intestinal bleeding.

Methods

In a cohort study of new users of nonsteroidal anti-inflammatory drugs (NSAIDs) from the Dutch Integrated Primary Care Information (IPCI) database, we identified all patients who experienced an upper gastrointestinal bleeding (UGIB). We used a large-scale regularized regression to fit two PS models using all structured and unstructured information in the EHR. We calculated hazard ratios (HRs) to estimate the risk of UGIB among selective cyclo-oxygenase-2 (COX-2) inhibitor users compared to nonselective NSAID (nsNSAID) users.

Results

The crude hazard ratio of UGIB for COX-2 inhibitors compared to nsNSAIDs was 0.50 (95% confidence interval 0.18-1.36). Matching only on age resulted in an HR of 0.36 (0.11-1.16), and of 0.35 (0.11-1.11) when further adjusted for sex. Matching on PS only, the first model yielded an HR of 0.42 (0.13-1.38), which reduced to 0.35 (0.96-1.25) when adjusted for age and sex. The second model resulted in an HR of 0.42 (0.13-1.39), which dropped to 0.31 (0.09-1.08) after adjustment for age and sex.

Conclusions

PS models can be created using unstructured information in EHRs. An incremental benefit was observed by matching on PS over traditional matching and adjustment for covariates.

INTRODUCTION

Electronic health records (EHRs) are primarily used for routine medical care, but secondary use of EHR data for observational research is becoming increasingly popular especially in studying of drug effects postmarketing [1]. In this era, data is used to generate information on drug safety and effectiveness in a cost-efficient way and by exploiting actual care patterns, which differ largely from experimental settings [2–5]. In an experimental setting such as in randomized clinical trials, the choice for a treatment is randomized, which would take care of potential confounding by indication [6]. In actual care, the treatment decision is usually influenced by measurable patient characteristics such as medical history, concomitant drug intake but also by personal prescriber preferences, which cannot be measured easily. This phenomenon of preferential prescribing is also known as channeling and may lead to confounding by indication [7,8]. A well-known example of channeling is the preference of doctors to prescribe selective cyclo-oxygenase-2 inhibitors (COX-2 inhibitors) over nonselective (ns) non-steroidal anti-inflammatory drugs (NSAIDs) to patients at risk of developing upper gastrointestinal bleeding (UGIB)[9,10], as the COX-2 inhibitors were developed on purpose to mitigate the GI effects of NSAIDs. Although clinical trials showed that COX-2 inhibitors are 'safer' than nsNSAIDs in relation to UGIB[11], observational studies showed no large differences between the rate of UGIB between COX-2 inhibitor and nsNSAIDs, possibly due to residual confounding by indications arising from channeling[12]. In order to obtain unbiased estimates in observational studies this confounding must be dealt with adequately. However, it is challenging to capture all relevant channeling factors in the EHR databases because information is not primarily recorded for research purposes. Moreover, relevant information may also be recorded in EHRs in an unstructured way [13,14].

Attempts to construct methods that deal with confounding have resulted in the propensity score method, the propensity score is an estimated conditional probability of receiving one particular treatment over another given a set of measured covariates [15], it can be regarded as a comprehensive way to look at channeling. Propensity score methods can be used to control for the unbalance between the treatment groups in order to estimate the comparative effectiveness of treatments [15]. Four different methods of using the propensity to reduce confounding have been described [16]: (1) matching on propensity score; (2) stratification on the propensity score; (3) inverse probability of treatment weighting using the propensity score; (4) and covariate adjustment using the propensity score. Typically, all variables related to either the outcome and/or exposure, are included in the propensity score model [17,18], sometimes these variables are not the exact confounding factors but proxies thereof [19]. Yet, identifying appropriate proxies in large EHRs is challenging. Schneeweiss *et al*. [20] proposed a high-dimensional propensity score (hd-PS) algorithm to empirically identify a large number of relevant covariates, with high prevalence, to control for confounding. In a case study on coxibs and NSAIDs using claims data in the USA, application of the hd-PS algorithm to control for confounding was found to produce an effect estimate for the risk of upper GI complications between coxibs vs. NSAIDs that was comparable to the one found in randomized trials[21]. The hd-PS model is constructed by using many covariates of which some could serve as proxies for unobserved factors that

otherwise may not be considered. Typically, only structured information such as diagnostic or procedure codes that is available in the claims databases, are included in the model. Rassen et al. [22] evaluated whether adding two-word phrases, present in patients' unstructured free-text data, to the propensity score model could improve validity of pharmacoepidemiology studies. Adjusting for two-word phrases resulted in an improvement in confounding adjustment. Electronic health records comprise much unstructured data and we propose that this information could also be used as proxies for potential confounding factors.

The aim of this study was therefore to assess whether unstructured text in EHRs can be used to construct a propensity score model that would allow to properly deal with confounding. We assessed the performance of propensity score models in addressing confounding by indication using as an example the association between selective COX-2 inhibitors and nonselective NSAIDs in relation to upper gastrointestinal bleeding.

METHODS

Data source

We used data from the Dutch Integrated Primary Care Information database (IPCI) [23], a population-based general practice EHR database. This database contains prospectively collected routine care data representing real-life practice. In the Netherlands, all citizens are registered with a general practitioner (GP), who acts as a gatekeeper to secondary and tertiary medical care. IPCI contains information on more than 1.8 million patients from 340 GP practices. For each individual person, all relevant medical information from primary and secondary care is documented in the medical record. Apart from patient demographics, the recorded information in the EHRs contain medical notes (including symptoms, physical examination, assessments and diagnoses), drug prescriptions, laboratory results, referrals for hospitalization or specialist care, and hospital discharge summaries. In the IPCI database, drug prescriptions are recoded according to the Anatomical Therapeutical Chemical (ATC) classification for research purposes [24]. Diagnoses are coded according to the International Classification for Primary Care (ICPC) [25]. Almost 60% of the medical records are clinical narratives, which do not contain coded information, but contain important information such as patient-reported symptoms and notes from the GP.

Selection of NSAID cohort

We created a cohort of all new adult (≥18 years) users of NSAIDs between 1996 and 2013. Patients had to be enrolled for at least one year in the database in order to be eligible for cohort entry. Within the NSAID cohort, we created episodes of 'new' NSAID use according to the following criteria: (a) at least six months of data available before NSAID exposure, (b) no prescription of any nonselective NSAID or selective COX-2 inhibitor in the previous six months (c) no mentioning of drug names, in the free-text, corresponding with NSAID-related ATC codes in the previous six months. The duration of a prescription was calculated by dividing the prescribed quantity by daily dose regimen. An NSAID episode continued when consecutive NSAID

prescriptions started before or within 30 days of the end of the duration of the previous prescription. The end of the episode was defined as the end of the last NSAID prescription (see figure 1). Episodes were classified as an nsNSAID or COX-2 inhibitor episode based on the first prescription in that episode being an nsNSAID or a COX-2 inhibitor, respectively. If a patient switched between exposure (from COX-2 inhibitor to nsNSAID or vice versa), the duration of the NSAID episode was ended at the switch of the exposure. A patient could have multiple NSAID episodes, but only if the above-mentioned criteria were met.

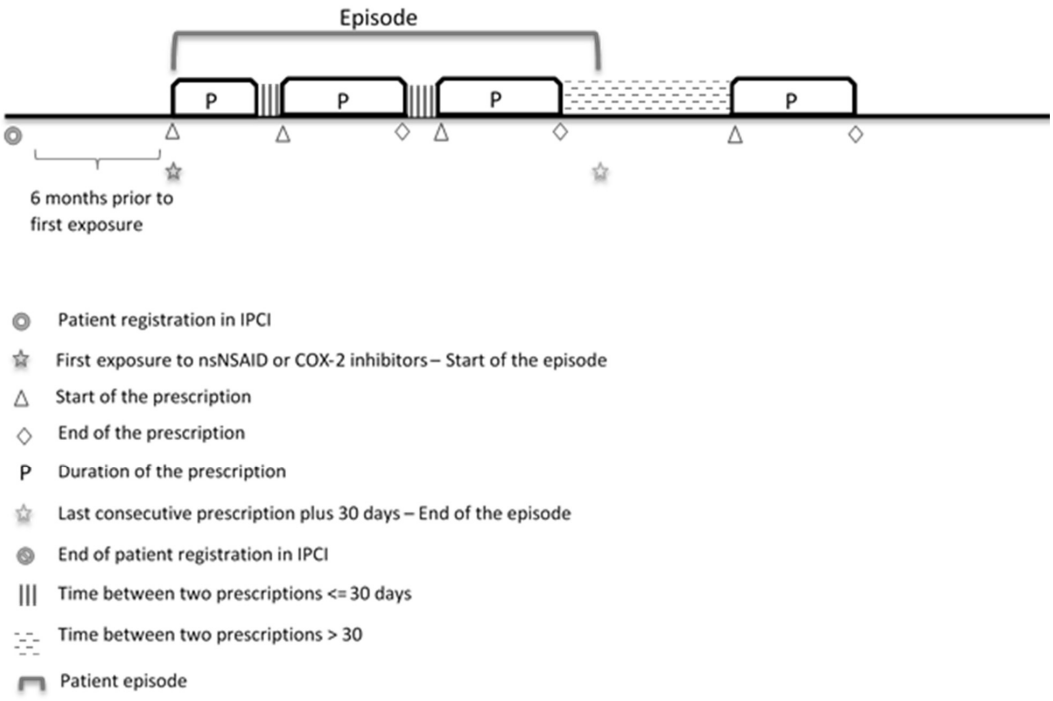


Figure 1: Episode selection

Selection of Upper Gastrointestinal bleeding patients

Within the cohort of new NSAID users, we identified all potential subjects who experienced an upper gastrointestinal bleeding (UGIB) via an automated search [26]. UGIB was defined as all forms of ulcer complications such as bleeding, perforation, or obstruction. The entire medical record of all potential UGIB patients was extensively reviewed to ensure the diagnosis and the date of onset. Any other cause of UGIB (such as variceal bleeding or Mallory Weiss bleeding) was excluded. The date of UGIB was determined as the date of first mentioning of symptoms leading to the UGIB diagnosis or if this date was unknown, the date of diagnosis.

Propensity score model

A propensity model was fitted using all information (structured and unstructured) in the EHR. To reduce the number of potential variables we first converted all text to lowercase after which we removed special characters, words not starting with a letter or a digit, stop words (such as *de*, *het* – the article *the* in English), and punctuation. All unique words (also known as unigrams) in the 6 months prior to cohort entry were extracted and used as textual features (potential covariates). This approach is commonly known as bag-of-words (BoW) model. We tested two methods to limit the number of covariates that would be included in the regression. The first method generated models using covariates of which the frequency in the cohort was above a certain threshold, e.g., 1000 without any further selection. In the second method, we generated a model using covariates that were associated with the outcome. The chi-square test was used to select covariates that were statistically significantly associated with the outcome (p-value less than 0.05). Another propensity model (method 3) was added for comparison, where only the established confounders (i.e. age, sex, and the exposure to low-dose aspirin) were included in the propensity score model. We used patients' prescription information to calculate exposure to low-dose aspirin.

The selected features were subsequently used in a large-scale regularized regression using a LaPlace prior [27] with the hyper-parameter of 0.01 to construct a propensity model for each method. The advantage of using a regularized regression is that it can handle high-dimensional data. A flowchart depicting the process of propensity score model generation (for methods 1 and 2) is presented in Figure 2.

Figure 2: Flowchart showing the process of generating a propensity score model from unstructured free-text

We used three-fold cross-validation [28] to evaluate the predictive accuracy of the models. The data set was randomly divided in three equally sized subsets or folds. In three cross-validation runs, each time, the model was successively trained on two folds and tested on the third fold. For each cross-validation run, an area under the receiver operating characteristic curve (AUC) was calculated. The averaged AUC was used as the overall performance measure.

One-to-many propensity score matching

The propensity score that was generated in each of the two models was used to account for the preferential prescribing of COX-2 inhibitors to patients at high-risk of developing an UGIB [12]. In this study, we used the greedy one-to-many matching as described by Rassen et al. [29]:

1. For each COX-2 inhibitor cohort member the difference in PS with each nsNSAID users was computed

2. Starting with the lowest difference, each COX-2 inhibitor cohort member was matched with one nsNSAID cohort member. Once an nsNSAID user was matched, he or she was precluded from further matching. A caliper of 0.01 was used, meaning no matches were made if the difference in PS was greater than 0.01.
3. After all COX-2 inhibitor cohort members were matched with one nsNSAID cohort member, the process was repeated until all nsNSAID users were matched or there was no match possible.

The algorithm ensured that all COX-2 inhibitor cohort members were matched with at least one nsNSAID cohort member if such a match was available within the caliper.

Statistical Analysis

To estimate the risk of UGIB among COX-2 inhibitor users compared to nsNSAID users we calculated hazard ratios with their corresponding 95% confidence intervals (CIs) using Cox proportional hazard regression. We conducted the analysis for four datasets: 1) a crude comparison (unmatched, no propensity score); 2) matched on age (± 2 years) and adjusted for sex and exposure to low-dose aspirin, no propensity score; 3) matched on PS with covariate frequency above 1000 and then adjusted for age, sex, and exposure to low-dose aspirin; and 4) matched on PS with covariates having an association with the outcome and then adjustment for age, sex, and exposure to low-dose aspirin.

RESULTS

NSAID cohort

From the source population of more than 1.8 million patients we identified 518,768 new users of NSAIDs based on ATC codes. We then processed the unstructured free-text in the entries of the new users to identify mentioning of drug names corresponding with NSAID-related ATC codes. In total, 36,188 new users were removed because either an nsNSAID or COX-2 inhibitor drug was mentioned in the free-text in the six months preceding first NSAID exposure. This resulted in 482,580 new NSAID users in the study cohort. Out of these, 459,701 (95%) were nsNSAID users and 22,879 (5%) were COX-2 inhibitor users.

Within the NSAID cohort, we retrieved 11,994 potential UGIB patients. After reviewing the medical records, we retained 1,048 UGIBs.

The average duration of episodes for initiators of COX-2 inhibitors was 94 days and 66 days for initiators of nsNSAIDs. Baseline characteristics of initiators of COX-2 inhibitors and nsNSAIDs are shown in Table 1. Most of the episodes of COX-2 inhibitors and nsNSAIDs were started after the year 2004.

Table 1: Baseline characteristics of initiators of selective COX-2 inhibitors or nsNSAIDs

Characteristics	**%**	
	COX-2 initiators (n=22,879)	**nsNSAID initiators (n=459,701)**
Age (mean)	57.7	47.9
Male	36.5	43.2
Female	63.5	56.8
Exposure to low-dose aspirin	2.8	1.1
Age (years)		
<= 30	6.5	17.3
31 – 40	8.4	16.1
41 – 50	17.7	22.4
51 – 60	22.4	19.7
61 – 70	20.8	13.8
71 – 80	15.9	7.7
> 80	8.3	3.0
Calendar year of treatment initiation		
before 2003	0.1	10.8
2003	1.4	2.0
2004	3.1	1.9
2005	1.6	1.9
2006	1.5	1.3
2007	2.6	2.3
2008	7.3	6.7
2009	11.5	12.3
2010	15.6	16.4
2011	22.7	20.6
2012	30.7	22.7
2013	1.9	1.1
UGI risk factors		
Use of antiplatelets	6.3	3.2
Use of anticoagulants	3.2	1.3
Use of gastroprotective agents	23.4	11.8
Other comorbidities		
Dyspepsia	0.2	0.2
Smoking	0.5	0.5
Heart failure	0.4	0.2
Diabetes mellitus	0.5	0.3
Concomitant use of other medication		
SSRIs	4.4	3.3
Spironolactone	0.7	0.3
Calcium channel blockers	7.2	3.7

Propensity model

In total, we extracted 2,762,326 covariates (i.e., unique words, out of almost 96 million words) from approximately 2.4 million entries in the 6 months prior to NSAID episodes from the medical records of 482,580 new NSAID users.

Table 2 shows the performance of the propensity models built using different covariates selection methods. The first model used all covariates with a frequency of 100 or more in the cohort, which resulted in 95,078 unique covariates entered into the model. Increasing the frequency to 1,000 resulted in a reduction of the number of covariates to 27,619. The number of covariates further reduced when frequency was increased to 5,000. The performance of the models in terms of their predictive accuracy was comparable. The predictive performance of the propensity model that was built using 3,650 covariates that had an association with the outcome according to the chi-square test. This resulted in an AUC of 70.59. The performance of the propensity model that included only the established confounders resulted in an AUC of 66.27. However, there were only 111 covariates in the model.

Table 2: Predictive performance of different propensity models

PS Model		Number of covariates	AUC*
Method 1	Covariate filtered on frequency ≥ 100	95,078	72.27
	Covariate filtered on frequency ≥ 1,000	27,619	72.32
	Covariate filtered on frequency ≥ 5,000	11,699	72.17
Method 2	Covariates filtered using Chi-square test (independent of frequency)	3,650	70.59
Method 3	Only established confounders (age, sex, and exposure to low-dose aspirin)	111	66.27

* AUC, area under the receiver operating characteristic curve

Risk of upper gastrointestinal bleeding

The crude hazard ratio of UGIB for COX-2 inhibitors compared to nsNSAIDs was 0.50 (95% 0.18-1.36) (Table 3). When matched on age, the hazard ratio of COX-2 inhibitor use compared to nsNSAID use was 0.36 (95% CI: 0.11-1.16). Further adjusting for sex and exposure to low-dose aspirin resulted in HR of 0.35 and 0.36 respectively. Matching on PS only, using one-to-many matching with a covariate frequency above 1,000, reduced the hazard ratio to 0.42 (95% CI: 0.13 – 1.38). Subsequent adjustment for age resulted in a hazard ratio of 0.36 (95% CI: 0.10 – 1.22). Matching on PS limiting to covariates that were associated to the outcome also provided a hazard ratio of 0.42 (95% CI: 0.13 – 1.39). Adjusting for age reduced the hazard ratio to 0.32 (95%: 0.09 – 1.09).

Table 3: Hazard ratios with 95% confidence intervals (CI) comparing COX-2 inhibitors with nsNSAIDs for different matching strategies and adjustments

Matching	Adjustment	Hazard ratio	95% CI
Unmatched	None	0.50	0.18 – 1.36
Age	None	0.36	0.11 – 1.16
	Sex	0.35	0.11 – 1.18
	Sex, Aspirin	0.36	0.11 – 1.18
Propensity Score (covariate frequencies >= 1000)	None	0.42	0.13 – 1.38
	Age	0.36	0.10 – 1.22
	Sex	0.39	0.12 – 1.30
	Age, Sex	0.35	0.16 – 1.25
	Sex, Aspirin	0.39	0.12 – 1.32
Propensity Score (covariates based on association test)	None	0.42	0.13—1.39
	Age	0.32	0.09—1.09
	Sex	0.43	0.13—1.42
	Age, Sex	0.31	0.09 – 1.08
	Sex, Aspirin	0.43	0.13—1.42
	Age, Sex, Aspirin	0.31	0.09 – 1.10

DISCUSSION

In this study, we generated a propensity model using unstructured information from EHRs. We tested different methods to construct this and demonstrated the feasibility to do so as well as its performance. Since electronic health records are now widely available for secondary use, we need to develop methods and test performance of these methods for use in epidemiological evaluations such as drug effects.

Our method to generate a propensity score model is substantially different from the high-dimensional propensity score (hd-PS) approach proposed by Schneeweiss et al [20]. The hd-PS algorithm that was developed for claims data uses structured information such as diagnostic codes, in-patient procedure codes, and drugs dispensed. In each identified data dimension, the highest ranked codes are selected to enter in the hd-PS model. The use of two-word free-text phrases in addition to the structured information has also been positively evaluated in the context of hd-PS models [22]. Our method is different since we used as the basis all unstructured text to generate propensity models, using a large-scale regularized regression, without pre-

identified data dimensions. Several methods other than logistic regression such as data-adaptive and classification trees have been proposed for fitting a propensity model [30]. To reduce the number of 'meaningless' features, we needed various textual data cleaning steps. We subsequently extracted all unigrams from the cleaned free-text, which served as potential covariates. Here we applied different approaches, to look at the impact of our choices. In the first method, the most-frequent covariates in the cohort were selected to enter the propensity score model. Since the covariates were selected merely based on their frequency in the cohort, this method is prone to include covariates that may actually be instrumental variables. Instrumental variables have an association with the exposure but not with the outcome except through their effect on exposure. If covariates are included that are not true confounders, the variance increases and sometimes a small amount of bias may be introduced [31–34]. In order to mitigate the potential to include covariates that are instrumental variables we included covariates with a significant association with the outcome to the propensity score model in the second method we applied [31].

We used three-fold cross-validation to evaluate the predictive performance of exposure to nsNSAID or Coxib for each generated PS model. In the first method where covariates were selected based on their frequency, increasing the frequency threshold for covariate selection reduced the number of covariates that entered into the propensity score model but the performance of the models was still comparable. This suggests that the performance of the models was mostly based on a few covariates with high occurrence in the text. Reducing the number of covariates reduced the computation time needed to fit the model. By selecting covariates with an association with the outcome, we significantly reduced the total number of covariates without greatly affecting the performance. The propensity models generated using covariates with only high frequency in the cohort performed better than the one where association with the outcome was verified. This may be due to the presence of some instrumental variables, which can result in an increase in predictive performance [30]. We used another propensity model for the comparison purposes where only the established confounders age, sex, and exposure to low-dose aspirin were included. The predictive performance of this model was lower than the other two models, which were generated from the free-text covariates. The second method, where covariate association with the outcome was verified, showed large decrease in the hazard ratios after further adjustments. Whereas previous studies have constructed the hd-PS with structured information, such as ICD and READ codes across different data dimensions in different sources [19–21,35], large proportions of information may be unstructured. We showed that this unstructured free-text can be used to construct propensity models. Initially, the new user cohort was created based on the prescription tables containing ATC codes. A high number of removals (7%) from the cohort based on the drug mentioned in the free-text indicate the importance of processing unstructured free-text instead of only relying on the structured information.

Our study also has several limitations. First, by including covariates based on their frequencies we might have selected covariates that are not necessarily related to the outcome or the exposure, which could introduce bias [18,36]. Second, since we only used unigrams, covariates like '*congestive heart failure*' cannot be recognized as such. Instead, it will be recognized as

individual words '*congestive'*, *'heart'*, and '*failure'*, which might lead to over- and underestimation of some covariates. Like previous studies using hd-PS methods, we also used the known association between NSAIDs and UGIB as an example. It is unclear whether our findings regarding the PS generated from unstructured free-text apply to other treatment-outcome pairs. Since the PS algorithm in general relies on the information present in the cohort, a similar approach using a different data set might have different results even when using known example of NSAID-UGIB.

The majority of COX-2 inhibitor episodes started after the year 2004, the period after the withdrawal of rofecoxib from the market because of cardiovascular risks [37]. This may explain the strong protective effect of COX-2 inhibitors in the crude analysis, which we would expect, but is different from previous observational studies that were done more closely to the introduction of coxibs [19–21,38]. Since most of our patients started after the contra-indications were introduced, channeling towards high risk patients was less of an issue [39].

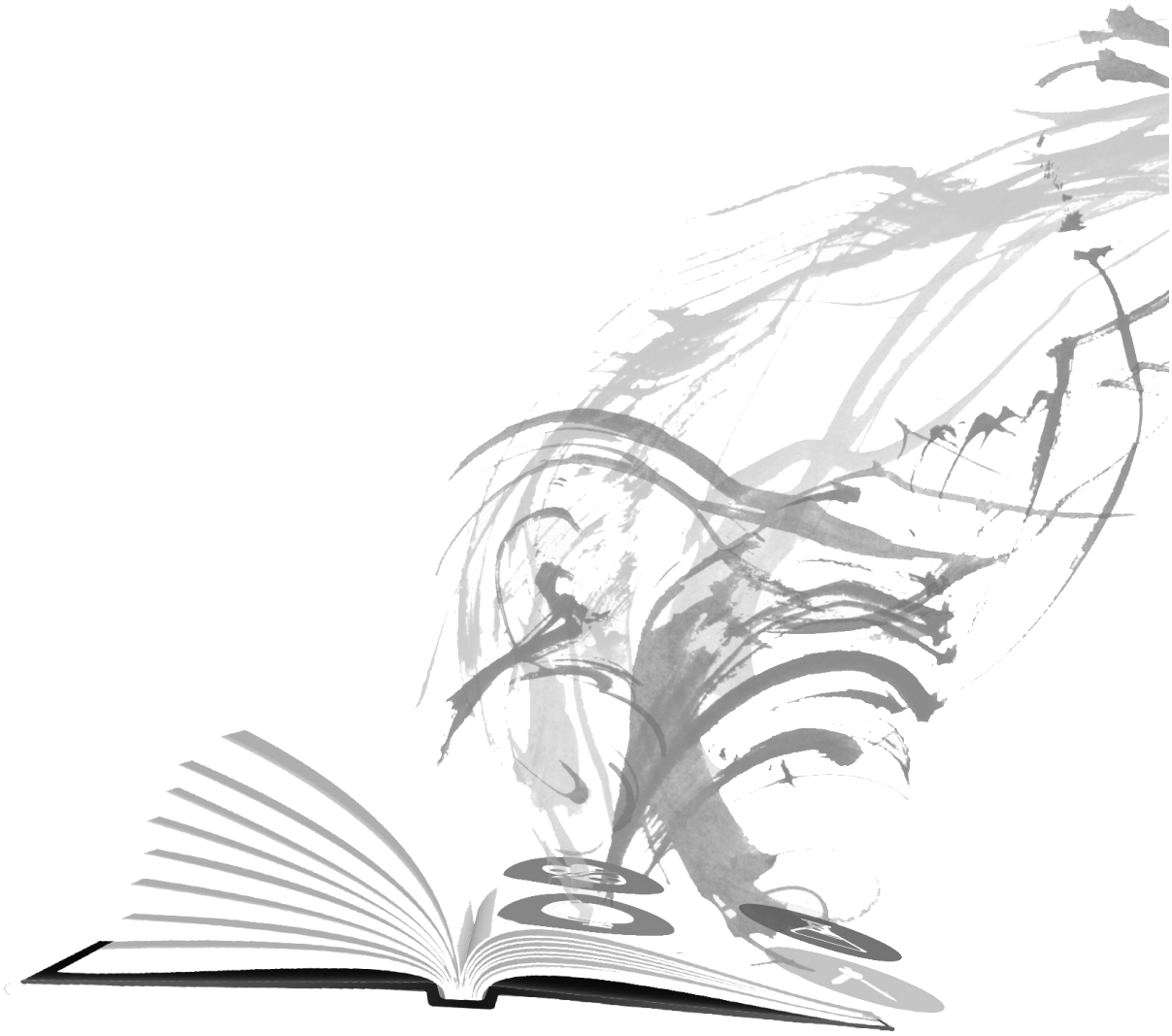
In conclusion, our study showed that PS models can be created using unstructured information in electronic healthcare records. We also showed that the PS model where covariates were filtered on their association with the outcome provide an improvement in adjustment for confounding. This is useful for database studies using a large amount of unstructured free-text as in EHRs. Better methods for extracting meaningful covariates from the free-text may be required for effective proxy adjustment via propensity scores.

REFERENCES

1. Strom B, Carson J: **Automated databases used for pharmacoepidemiology research**. *Clin Pharmacol Ther* 1989, **46**:390–394.
2. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW: **Secondary use of electronic health record data: Spontaneous triggered adverse drug event reporting**. *Pharmacoepidemiol Drug Saf* 2010, **19**:1211–1215.
3. Allen-Dicker J, Klompas M: **Comparison of electronic laboratory reports, administrative claims, and electronic health record data for acute viral hepatitis surveillance.** *J Public Health Manag Pract* 2012, **18**:209–214.
4. Schneeweiss S, Avorn J: **A review of uses of health care utilization databases for epidemiologic research on therapeutics.** *J Clin Epidemiol* 2005, **58**:323–337.
5. Suissa S, Garbe E: **Primer: administrative health databases in observational studies of drug effects--advantages and disadvantages.** *Nat Clin Pract Rheumatol* 2007, **3**:725–732.
6. Salas M, Hofman A, Stricker BH: **Confounding by indication: An example of variation in the use of epidemiologic terminology**. *Am J Epidemiol* 1999, **149**:981–983.
7. Walker AM: **Confounding by indication.** *Epidemiology* 1996, **7**:335–336.
8. Mosis G, Stijnen T, Castellsague J, Dieleman JP, van der Lei J, Stricker BHC, Sturkenboom MCJM: **Channeling and prevalence of cardiovascular contraindications in users of cyclooxygenase 2 selective nonsteroidal antiinflammatory drugs.** *Arthritis Rheum* 2006, **55**:537–542.
9. Masclee GM, Valkhoff VE, Coloma PM, de Ridder M, Romio S, Schuemie MJ, Herings R, Gini R, Mazzaglia G, Picelli G, Scotti L, Pedersen L, Kuipers EJ, van der Lei J, Sturkenboom M: **Risk for upper gastrointestinal bleeding from different drug combinations.** *Gastroenterology* 2014.
10. Moride Y, Ducruet T, Boivin J-F, Moore N, Perreault S, Zhao S: **Prescription channeling of COX-2 inhibitors and traditional nonselective nonsteroidal anti-inflammatory drugs: a population-based case-control study.** *Arthritis Res Ther* 2005, **7**:R333-42.
11. Bhala N, Emberson J, Merhi A, Abramson S, Arber N, Baron JA, Bombardier C, Cannon C, Farkouh ME, FitzGerald GA, Goss P, Halls H, Hawk E, Hawkey C, Hennekens C, Hochberg M, Holland LE, Kearney PM, Laine L, Lanas A, Lance P, Laupacis A, Oates J, Patrono C, Schnitzer TJ, Solomon S, Tugwell P, Wilson K, Wittes J, Baigent C: **Vascular and upper gastrointestinal effects of non-steroidal anti-inflammatory drugs: meta-analyses of individual participant data from randomised trials.** *Lancet* 2013, **382**:769–779.
12. van Soest EM, Valkhoff VE, Mazzaglia G, Schade R, Molokhia M, Goldstein JL, Hernández-Díaz S, Trifirò G, Dieleman JP, Kuipers EJ, Sturkenboom MCJM: **Suboptimal gastroprotective coverage of NSAID use and the risk of upper gastrointestinal bleeding and ulcers: an observational study using three European databases.** *Gut* 2011, **60**:1650–9.
13. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA: **Extracting information from the text of**

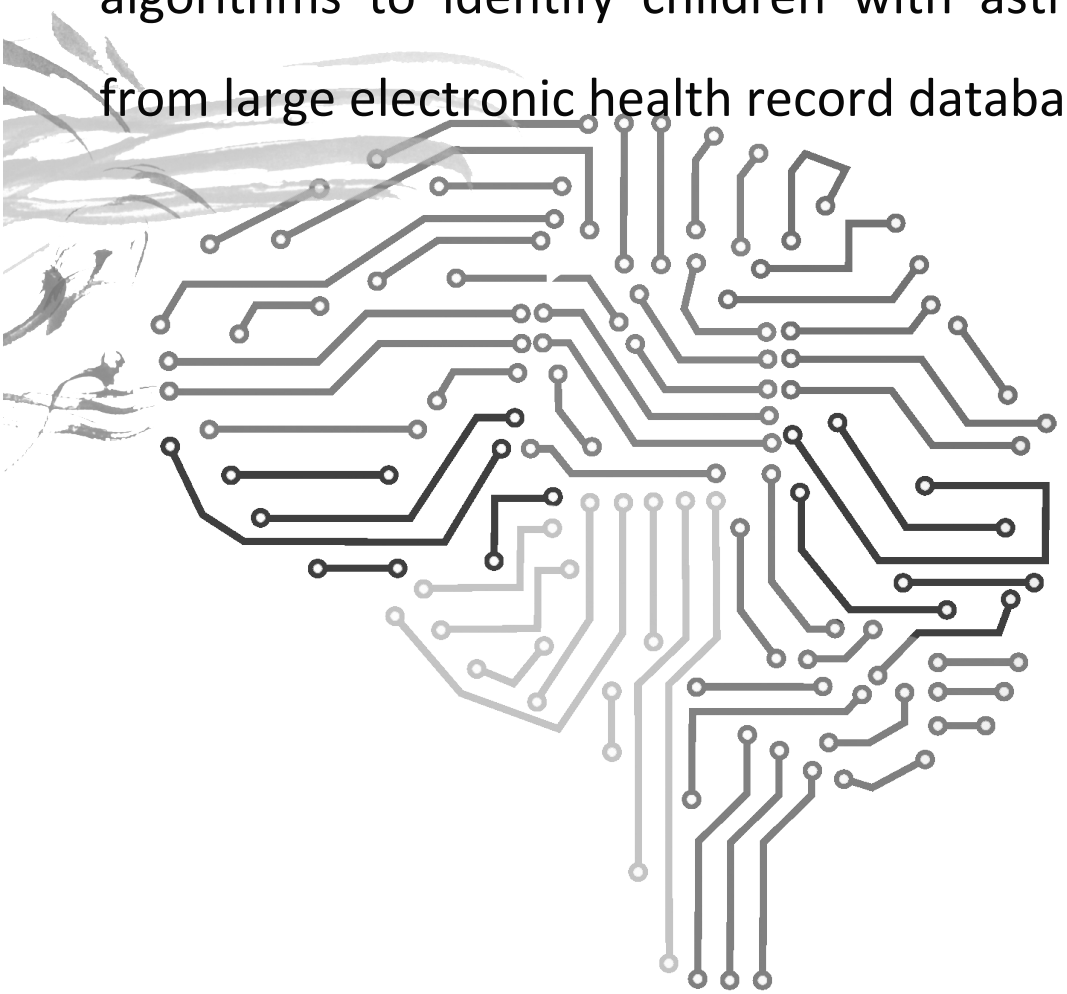
electronic medical records to improve case detection: a systematic review. *J Am Med Informatics Assoc* 2016:ocv180.
14. Yadav P, Steinbach M, Kumar V, Simon G: **Mining electronic health records: A survey**. *ACM Comput Surv* 2016, **1**:1–41.
15. Rosenbaum P, Rubin D: **The central role of the propensity score in observational studies for causal effects**. *Biometrika* 1983, **70**:41–55.
16. Austin PC: **An introduction to propensity score methods for reducing the effects of confounding in observational studies.** *Multivariate Behav Res* 2011, **46**:399–424.
17. Rubin DB: **Estimating causal effects from large data sets using propensity scores.** *Ann Intern Med* 1997, **127**:757–763.
18. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T: **Variable selection for propensity score models**. *Am J Epidemiol* 2006, **163**:1149–1156.
19. Toh S, Rodriguez LA, Hernan MA: **Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records**. *Pharmacoepidemiol Drug Saf* 2011, **20**(June):849–857.
20. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA: **High-dimensional propensity score adjustment in studies of treatment effects using health care claims data.** *Epidemiology* 2009, **20**:512–522.
21. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S: **High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications.** *Eur J Clin Pharmacol* 2013, **69**:549–557.
22. Rassen JA, Wahl PM, Angelino E, Seltzer MI, Rosenman MD: **Automated Use of Electronic Health Record Text Data To Improve Validity in Pharmacoepidemiology Studies**. In *Pharmacoepidemiol Drug Saf*. *Volume 22*. NJ USA: WILEY-BLACKWELL; 2013:376.
23. Vlug A, van der Lei J, Mosseveld B, van Wijk M, van der Linden P, MC S, van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project**. *Methods Inf Med* 1999, **38**:339–344.
24. **WHO Collaborating Centre for Drug Statistics Methodology. Guidelines for ATC classification and DDD assignment** [http://www.whocc.no/atcddd/]
25. Lamberts H, Wood M: **ICPC: International Classification of Primary Care**. *Scand J Prim Heal Care* 1987:204.
26. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors J a: **Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records.** *BMC Med Inform Decis Mak* 2013, **13**:30.
27. Genkin A, Lewis DD, Madigan D: **Large-scale bayesian logistic regression for Text categorization**. *Technometrics* 2007, **49**:291–304.
28. Mosteller F, Tukey JW: **Data Analysis, Including Statistics**. In *Handb Soc Psychol Vol 2 Res methods*. Edited by Lindzey G, Aronson E. Reading, MA: Addison-Wesley; 1968:80–203.
29. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S: **One-to-many propensity score matching in cohort studies**. 2012, **21**:69–80.

30. Pirracchio R, Petersen ML, van der Laan M: **Improving propensity score estimators' robustness to model misspecification using super learner**. *Am J Epidemiol* 2015, **181**:108–119.
31. Myers J a., Rassen J a., Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ: **Effects of adjusting for instrumental variables on bias and precision of effect estimates**. *Am J Epidemiol* 2011, **174**:1213–1222.
32. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S: **Confounding control in healthcare database research: challenges and potential approaches**. *Med Care* 2010, **48**:S114–S120.
33. Rassen JA, Schneeweiss S: **Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system**. *Pharmacoepidemiol Drug Saf* 2012, **21**:41–49.
34. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S: **Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database Analyses**. *Am J Epidemiol* 2015, **182**:651–659.
35. Le H V, Poole C, Brookhart MA, Schoenbach VJ, Beach KJ, Layton JB, Stürmer T: **Effects of aggregation of drug and diagnostic codes on the performance of the high-dimensional propensity score algorithm: an empirical example.** *BMC Med Res Methodol* 2013.
36. Hernan MA, Hernandez-Diaz S, Robins JM: **A structural approach to selection bias**. *Epidemiology* 2004, **15**:615–625.
37. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Day R, Ferraz MB, Hawkey CJ, Hochberg MC, Kvien TK, Schnitzer TJ: *Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis. VIGOR Study Group. Volume 343*; 2000.
38. Le H V, Poole C, Brookhart MA, Schoenbach VJ, Beach KJ, Layton JB, Stürmer T: **Effects of aggregation of drug and diagnostic codes on the performance of the high-dimensional propensity score algorithm: an empirical example.** *BMC Med Res Methodol* 2013, **13**:142.
39. Sun SX, Lee KY, Bertram CT, Goldstein JL: **Withdrawal of COX-2 selective inhibitors rofecoxib and valdecoxib: Impact on NSAID and gastroprotective drug prescribing and utilization.** *Curr Med Res Opin* 2007, **23**:1859–1866.



Chapter 5

Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases



ABSTRACT

Purpose

Most electronic health record (EHR) databases contain unstructured free-text narratives, which cannot be easily analyzed. Case detection algorithms are usually created manually and often rely only on using coded information such as ICD-9 codes. We applied a machine-learning approach to generate and evaluate an automated case detection algorithm that uses both free-text and coded information to identify asthma cases.

Methods

The Integrated Primary Care Information (IPCI) database was searched for potential asthma patients aged 5-18 years using a broad query on asthma-related codes, drugs, and free-text. A training set of 5032 patients was created by manually annotating the potential patients as definite, probable, or doubtful asthma cases, or non-asthma cases. The rule-learning program RIPPER was then used to generate algorithms to distinguish cases from non-cases. An over-sampling method was used to tune the performance of the automated algorithm to meet our study requirements. Performance of the automated algorithm was evaluated against the manually annotated set.

Results

The algorithm yielded a positive predictive value (PPV) of 0.66, sensitivity of 0.98 and specificity of 0.95 when identifying only definite asthma cases, a PPV of 0.82, sensitivity of 0.96, and specificity of 0.90 when identifying both definite and probable asthma cases, and a PPV of 0.57, sensitivity of 0.95, and specificity of 0.67 for the scenario identifying definite, probable, and doubtful asthma cases.

Conclusions

The automated algorithm shows good performance in detecting asthma cases utilizing both free-text and coded data. This algorithm will facilitate large-scale studies of asthma in the IPCI database.

INTRODUCTION

Asthma is one of the most common chronic diseases of childhood globally. The main goal of asthma treatment is to achieve and maintain clinical control of the disease. Failing to control asthma can limit daily-life activities and can be fatal. In children, asthma is usually treated and maintained with low-dose inhaled corticosteroids (ICS). If asthma is not controlled, treatment is stepped up by either adding long acting B_2 agonists (LABA) or a leukotriene receptor antagonist (LTRA) to low-dose ICS or increasing the dose of ICS until control is achieved [1].

Safety concerns have been raised on the long-term toxicity of ICS, the risk of mortality, and asthma exacerbations with the use of LABAs in monotherapy and the risk of neuropsychiatric events and hepatotoxicity in children treated with LTRAs [2–8]. Randomized controlled trials (RCTs) on the efficacy and safety of these drugs in children are rare. In addition, the few trials conducted in children are often not designed to detect safety issues because of the limited sample size and short duration of follow-up. In general, observational studies are suited for research on drug safety because they usually have large sample size with long-term follow-up. Electronic medical records are valuable resources and are increasingly being used in epidemiological observational studies to detect safety issues [9–15].

One of the challenges of using electronic medical records is to determine whether and when a medical outcome of interest has occurred. When coded information such as International Classification of Diseases version 9 (ICD-9) and Logical Observation Identifiers Names and Codes (LOINC) codes are available, outcomes are typically identified by searching for a combination of codes in the patient record. However, the recording of these codes can be incomplete and inaccurate, or the codes themselves might be ambiguous or have the wrong granularity for the research question at hand. It is therefore recommended that the performance of this search using codes is evaluated through manual chart review, where researchers often rely on the free-text narrative in the medical record. There are also databases where the coding is simply too incomplete. For example, in the Integrated Primary Care Information (IPCI) database [16], almost 60% of the record lines are narratives and do not contain coded information. These narratives may contain important information such as patient-reported symptoms, signs, or summaries of specialists' letters. In these databases, the search for outcomes is even more labor intensive. Usually, a broad text query is defined including all possible words and codes that might be relevant, and subsequently all narratives returned by the query are manually reviewed. With the increase in size of these databases, this practice is becoming prohibitively laborsome and expensive.

For this reason, we used an alternative approach to identifying asthma cases, which uses the free-text narrative in an automated fashion. We apply a machine-learning approach to derive an automated case detection algorithm that uses both text and coded data if available. We not only show the performance of this algorithm in terms of positive predictive value (PPV), sensitivity, and specificity, but also demonstrate how sensitivity and specificity can be tuned to best meet the requirements of our study. We apply this approach to the Dutch IPCI database, but the same

procedure to construct a case detection algorithm could be used on other databases, in other languages.

METHODS

Electronic medical record database

Data in this study were taken from the IPCI database [16]. The IPCI database is a longitudinal observational database of electronic medical records (EMRs) from Dutch general practitioners (GPs). The electronic records contain coded data and data on patient demographics, symptoms and diagnoses, clinical findings, referrals, laboratory findings, and hospitalization of more than 1.1 million patients. The IPCI database uses the International Classification of Primary Care (ICPC) [17] coding system. The cohort for the underlying study included children between age 5 and 18 that were present in the database between the dates January 1, 2000 until January 31, 2012. A minimum registration period of six months was required to guarantee sufficient medical history data.

Clinical case definition

To create a labeled training set for machine-learning methods, we used a manually defined clinical case definition. Patients were categorized into 'definite asthma', 'probable asthma', 'doubtful asthma', or 'no asthma' according to the following validation criteria.

For definite asthma patients, at least one entry in their medical record containing an asthma diagnosis confirmed by a specialist (pediatrician or pulmonologist) was required. For probable asthma patients, at least one entry should contain evidence of asthma diagnosed by the GP and there should be at least one more entry in the patient record suggestive of asthma (ICPC code, free text, lung function measurements, or use of specific bronchodilating drugs/anti-inflammatory drugs for the indication of asthma) within the next 12 months, or at least two additional entries in the patient record suggestive of asthma. For doubtful asthma patients, there should be at least one entry containing an indication of asthma without satisfying the criteria for a definite or probable asthma case. Patients with drug entries only (i.e., without evidence in ICPC code or free text) were considered non-asthma cases, as were patients with no indication of asthma in any entry of their patient record.

Training set for machine learning

We use machine-learning methods to automatically learn case detection algorithms based on a training set of entries, i.e., a set of positive and negative examples. To generate a training set for our machine-learning method, we first identified all potential asthma patients using a broad automated search on ICPC asthma disease codes, asthma drug prescriptions, and free text. Because of the generic asthma-related keywords used in the broad query, many of the retrieved patients were likely not to have asthma.

One medical doctor reviewed the entire medical record of the patients identified by the broad search strategy in random order for one month. A total of 5032 patients were validated from 63,618 patients returned by the broad query. A senior medical doctor further reviewed the doubtful patients. Patients were labeled as definite asthma (n=308), probable asthma (n=1133), doubtful asthma (n=160), or non-asthma (n=3431). A patient's medical record consists of one or more entries, where each entry pertains to a patient visit, a letter from a specialist, prescribed drugs, and so on. The entries in the medical record of a patient were reviewed in chronological order and a patient was labeled positive whenever an asthma criterion (for definite, probable, and doubtful) was satisfied. The remaining entries in the medical record were not reviewed, and only the entries containing the indication of asthma were included in the training set as positive examples. If none of the entries of the patient record contained positive evidence of asthma, the patient was considered a negative case and one of the entries was randomly picked as a negative example in the training set.

To make the text in the entries better fit for machine learning approaches, we removed uninformative words (so-called stop words). Although some standard Dutch stop word lists are available [18], they are not entirely suitable for the clinical text because some of the words may have importance in the clinical context, e.g., 'op' (English 'on') could be an abbreviation of 'operation'. We therefore used a small stop word list, only containing 'en' (English 'and'), 'een' ('a'), 'de', and 'het' (both 'the').

All ATC (Anatomical Therapeutic Chemical) codes related to respiratory drugs and starting with R03, (drugs for obstructive airway diseases) were replaced by a single keyword 'r03drug'. To remove negated and speculative assertions, we used the Dutch assertion filter proposed in [15], similar to NegEx [19]. Any words appearing between negation or speculation keywords and the end of sentence (demarked by a punctuation mark) were removed from the entry. All sentences containing an alternatives keyword were completely removed.

The text in the entries was converted to lower case and split into individual words. These individual words were treated as features for our machine-learning method (bag-of-words representation). Schuemie et al. [15] previously showed the advantage of using assertion filtering and bag-of-words representation on Dutch EMRs. For computational purposes, we reduced the number of features by chi-square feature selection [20]. A p-value of less than 0.05 was used as a feature selection criterion. Chi-square feature selection decreased the number of features in the data set by about a factor of 10 without affecting the performance of the classifiers but greatly reducing their training time.

Automated generation of case definitions

Considering the hierarchical nature of the asthma labels (definite->probable->doubtful->non-asthma), we tackled the automated generation of case definitions as a hierarchical multi-class classification problem [21–23]. We followed an approach in which the hierarchy is structured as a decision tree and separate classifiers are built for the nodes in the tree (Figure. 1).

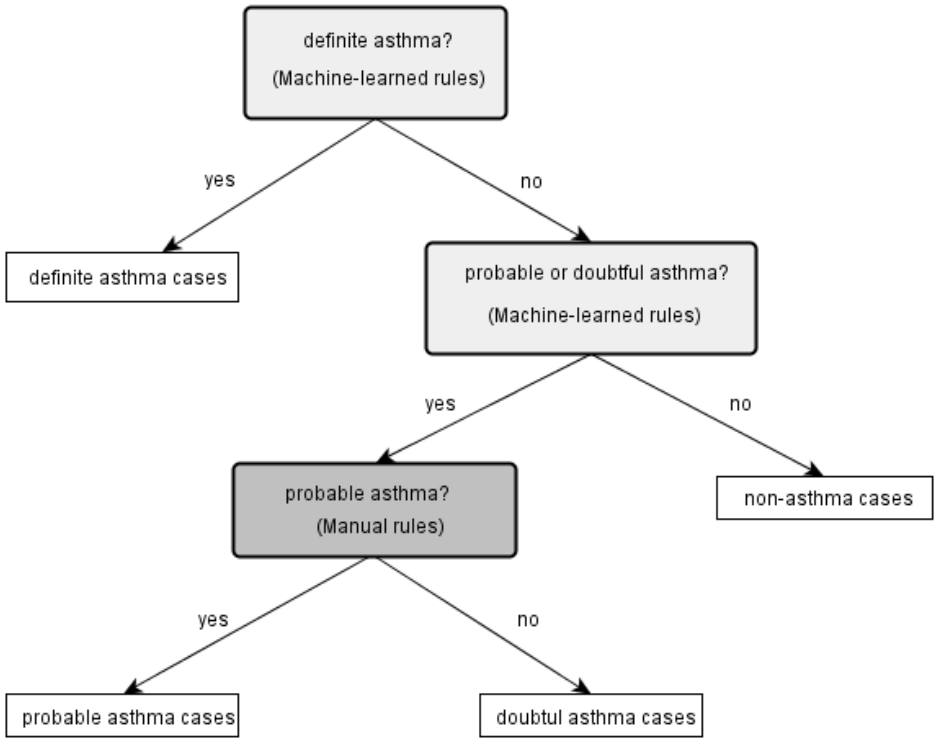


Figure 1: Hierarchical classification scheme for asthma

We trained two machine-learning classifiers, one to separate definite cases from all other cases and the other to distinguish probable and doubtful asthma from non-asthma cases. The second classifier considered probable and doubtful cases as one (positive) group because the distinction between these cases is difficult to learn automatically. This distinction was made in a third classifier, which implemented two rules based on the manual case definition criteria: (1) if a patient had two positive asthma entries (according to the second classifier) within a period of 15 months and (2) one of the entries is not a medication/drug entry, the patient was considered a probable asthma case. A medication entry only contains prescription. The training set for the first, 'definite asthma' classifier consisted of the entries of the definite asthma patients as positive examples, and entries of the probable, doubtful, and non-asthma cases as negative examples. For the second, 'probable, doubtful, and non-asthma' classifier, we used the entries of the probable and doubtful cases as positive examples and the entries of the non-asthma cases as negative examples. The probable or doubtful asthma patients classified as definite asthma by the first classifier were removed from the training set of the second classifier, and the definite asthma patients missed by the first classifier were added as positive examples.

To shift the balance of sensitivity and specificity, we used a method called "over-sampling". Several over-sampling methods are described in [24]. Normally over-sampling is done by simply reusing the same examples multiple times, but Schuemie et al. [15] showed that using additional entries of non-cases could lead to an increased performance. Our over-sampling was focused on increasing the specificity of the classifiers. For a non-asthma patient, all entries were manually reviewed and no evidence of asthma was found. Although initially one entry was randomly selected for training the classifiers, the other entries can also be used as additional negative examples. We created a set of all these additional entries, and randomly sampled from this set to expand our training set. In total, we used 10 over-sampling percentages in the experiments. In each over-sampling run, a specified percentage of entries (of the original negative examples in the training set) from the additional entries set were added to the training set. The experiment without the over-sampling entries was considered a baseline.

Training and testing

The rule-learning algorithm RIPPER [25] was used on the training set to automatically generate rules for each of the asthma case definition. Schuemie et al. [15] evaluated several well-known machine-learning algorithms for the classification of EMRs in a similar experimental setting, and found RIPPER to be one of the best performing algorithms. RIPPER produces an ordered set of decision rules. The advantages of such machine-learning algorithms are their ability to produce output that is understandable by humans, their ease of use, and their applicability to a wide range of problems [26]. We used an implementation of the RIPPER algorithm called JRip, which is available in the open-source machine-learning package Weka [27].

We used five-fold cross-validation to evaluate our classifiers. Cross validation was done at the patient level (subject-level cross-validation [28]) i.e., the data set was randomly divided in five equally sized subsets of patients (folds). In five cross-validation runs, each time the entries pertaining to four folds were used as a training set and the entries of the remaining subset were used for testing. We used all entries of the patients in the test fold because in real-life situations we do not know the labels of the entries pertaining to the patients returned by the broad query. Cross-validation was used to obtain unbiased performance estimates of the classifiers, but all data was used to generate the final sets of rules.

We used PPV, sensitivity, and specificity as measures to evaluate the performance of the classifiers. PPV is defined as the fraction of positively identified cases that are true positive: number of true positives / (number of true positives + number of false positives). Sensitivity is defined as the true-positive rate: number of true positives / (number of true positives + number of false negatives), whereas specificity is the true-negative rate: number of true negatives / (number of true negatives + number of false positives).

RESULTS

We present results of the asthma classification using three different scenarios. Each can be used to meet the requirement of a particular study. In the first scenario (Figure 2), only definite asthma cases were considered relevant for the study. The probable and doubtful asthma cases were

ignored. In the second scenario (Figure 3), the definite and probable asthma cases were considered relevant for the study. The definite and probable asthma cases were combined as positive asthma cases and doubtful cases were disregarded. In the third scenario (Figure 4), the definite, probable, and doubtful asthma cases were combined as positive asthma cases.

The sensitivity, specificity, and PPV of the classifiers using over-sampling and cross-validation methods for the three scenarios are presented in Figures 2-4.

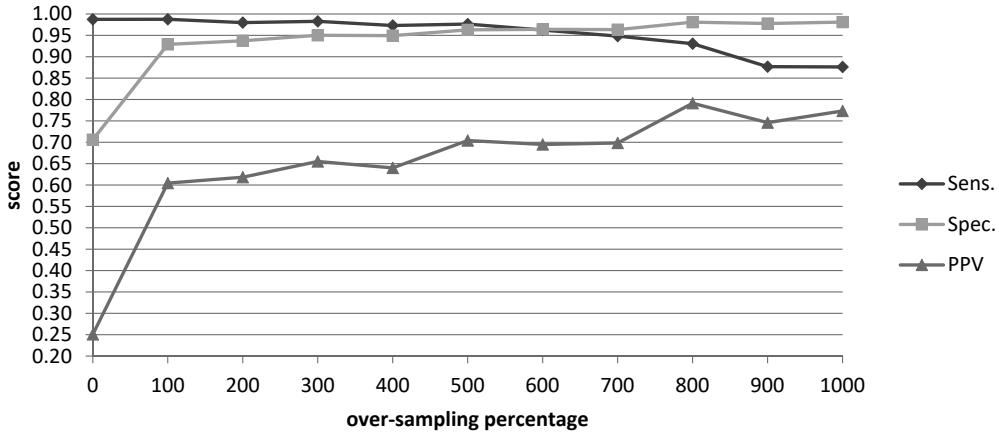


Figure 2: Performance of the classifiers using cross-validation when only definite asthma cases were considered as positive asthma ignoring probable and doubtful cases

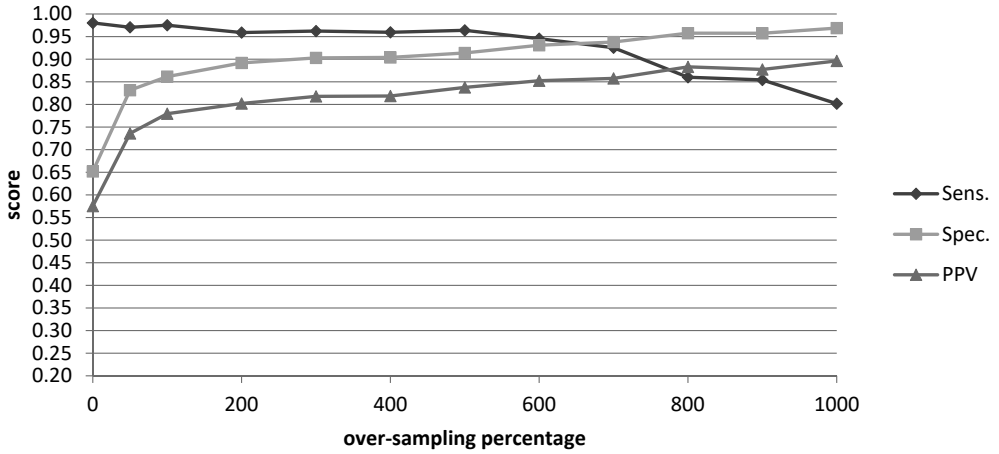


Figure 3: Performance of the classifiers using cross-validation when definite and probable asthma cases were combined as positive asthma ignoring doubtful cases

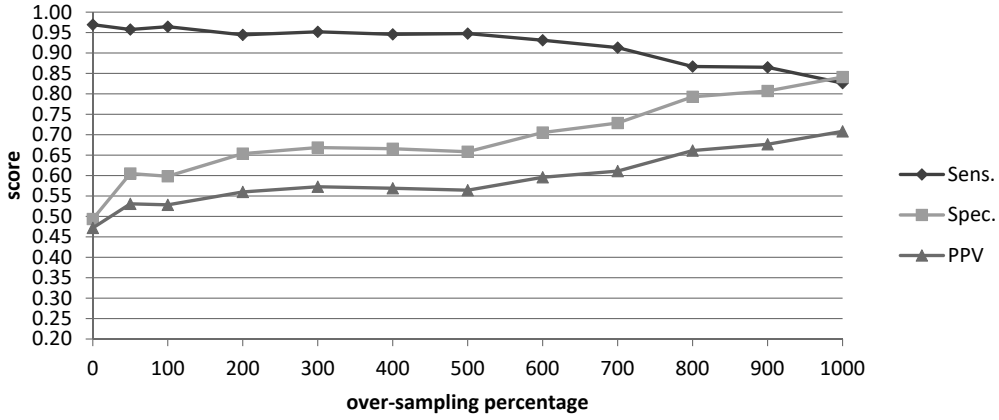


Figure 4: Performance of the classifiers using cross-validation when definite, probable, and doubtful asthma cases were combined as positive asthma

The first experiment with 0% over-sampling was considered as the baseline in our experiments. The classifiers showed consistent behavior during the over-sampling experiments. The specificity and PPV gradually increased and sensitivity decreased as we increased the number of negative

examples in the training set. For this particular study, we selected the model using 300% over-sampling as the final classification model because of its high sensitivity and specificity. A confusion matrix of the selected classification model is presented in Table 1.

Table 1: Confusion matrix of the case detection algorithm generated with 300% over-sampling using cross-validation

		Case detection algorithm				
		Definite asthma	Probable asthma	Doubtful asthma	Non-asthma	Total patients
Gold standard	Definite asthma	228	47	29	4	308
	Probable asthma	166	682	245	40	1133
	Doubtful asthma	16	15	96	33	160
	Non-asthma	120	130	887	2294	3431
	Total classified	530	874	1257	2371	5032

From 1601 asthma cases (definite, probable, and doubtful), only 77 (5%) were misclassified as non-asthma cases. From 3431 non-asthma cases, 1137 (33%) were misclassified as asthma cases. The automatic case definition for definite asthma is shown in Table 2 and for probable and doubtful asthma in Table 3. Where necessary, the English translation of the terms is included between parentheses.

Table 2: Automatically generated case detection rules for definite asthma

1. "20" and "astma" → true
2. "cmi" and "astma" and not "00" and not "van" (“from”) and not "s" → true
3. "cmi" and "flixotide" and not "ventolin" and not "medicatie" (“medicine”) and "20" → true
4. "cmi" and "kindergeneeskunde" (“pediatrician”) and "ventolin" and not "te" (“to”) → true
5. "cmi" and "astma" and "15" → true
6. "cmi" and "20" and "pulmicort" → true
7. "cmi" and "longziekten" (“pulmonologist”) → true
8. DEFAULT → false

Table 3: Automatic case definition for probable and doubtful asthma

1. "r03drug" and "r96" and not "01" → true
2. "r03drug" and not "hoesten" ("coughing") and not "pulm" and "flixotide" → true
3. "r03drug" and not "hoesten" and not "pulm" and not "piepende" ("wheezing") and not "hoest" ("cough") → true
4. "astma" and "r96" → true
5. "r03drug" and not "pulm" and "inh" → true
6. "ventolin" and not "pulm" and "r96" → true
7. "ventolin" and "astma" and not "vag" → true
8. "r03drug" and not "piepen" ("wheeze") and not "hoest" and "diskus" → true
9. DEFAULT → false

Table 4: Performance of case detection algorithms that were generated using different combinations of information present in the electronic medical records

	Scenario 1			**Scenario 2**			**Scenario 3**		
Information	Sens	Spec	PPV	Sens	Spec	PPV	Sens	Spec	PPV
Codes	0.53	0.87	0.21	0.56	0.85	0.57	0.62	0.76	0.55
Codes+Medications	0.86	0.67	0.18	0.67	0.67	0.42	0.69	0.60	0.45
Free text	0.88	0.96	0.64	0.62	0.94	0.78	0.68	0.81	0.63
Free text+Codes	0.85	0.95	0.62	0.61	**0.94**	0.77	0.65	**0.84**	**0.66**
Free text+Medications	0.84	**0.97**	**0.68**	0.62	0.94	0.79	0.68	0.81	0.63
Free text+Codes+Medications	**0.98**	0.95	0.66	**0.96**	0.90	**0.82**	**0.95**	0.67	0.57

The term 'cmi' indicates an incoming communication (i.e., a letter) from a specialist or outpatient GP. There are codes to identify specialties in IPCI and the numbers '20' and '15' are used for pediatrics and pulmonology, respectively. Because specialists do not code events in their communications with GPs, none of the rules contains an ICPC or ATC code. The drugs 'flixotide', 'ventolin', and 'pulmicort' are used for obstructive airway diseases. The terms 'r96', '00' and '01' are part of the asthma ICPC codes 'R96.00' and 'R96.01'. Our preprocessing algorithm separated the codes as 'R96', '00', and '01' and because of the bag-of-words representation, these were treated as individual features. The term 's' is part of the 'SOEP' registration used by the GPs in the Netherlands. The 'S' in 'SOEP' stands for 'subjective', and refers to patient history and symptoms. Since the SOEP and ICPC codes can be entered by the GPs only, entries containing

these terms indicate that these are GP entries. The keyword 'r03drug' marks the presence of an ATC code starting with R03, indicating a respiratory drug. The terms 'pulm', 'inh', and 'vag' are short for 'pulmonary', 'inhaler', and 'vesiculair ademgeruis' ('vesicular breath sound)', respectively. The term 'diskus' indicates a type of dry powder inhaler. For the words 'van' (English 'from' or 'of') and 'te' (English 'too') we have no reasonable explanation why RIPPER found them useful. Almost all rules for probable and doubtful asthma classification contain a mixture of codes and free text.

To assess the impact of different types of information (codes, medications, free text) on classification performance, we compared the performance of our selected model (using 300% over-sampling), generated using all information in the medical records, with models that were generated using subsets of information (also using 300% over-sampling). The results in Table 4 show that the models that only used codes or codes and medications have much lower performance than the models that use free text. None of the models comes close to our selected model with regard to sensitivity, while specificity and PPV of the reference model is comparable to those of the other models using free text for scenarios 1 and 2.

DISCUSSION

We created and evaluated an automated case detection algorithm to identify children with asthma within the IPCI database. The case detection algorithm was generated using a rule-learning algorithm which incorporated both information contained in the unstructured free-text and coded data in electronic medical records. We evaluated the automated algorithm in the context of three scenarios, and each scenario had different performance characteristics suitable for a different asthma study goal.

By using over-sampling techniques, we could vary the performance of the resulting detection algorithm. By adding more negative examples of asthma cases, PPV and specificity increased, at the cost of decreased sensitivity (cf. Figures 2-4). Varying the amount of over-sampling allows researchers to generate a case detection algorithm suitable for a specific study. For example, when investigating incidence and prevalence, where the goal is to find the number of cases in a population in a given time period, a case detection algorithm with high sensitivity would be preferred. For our particular asthma study, we selected the algorithm with 300% over-sampling mainly because of both its high specificity and sensitivity. The selected case detection algorithm had a PPV of 0.66, sensitivity of 0.98 and specificity of 0.95 for the scenario when only definite cases were considered relevant for the study (cf. Figure 2), PPV of 0.82, sensitivity of 0.96, and specificity of 0.90 for the scenario when definite and probable asthma cases combined were considered relevant (cf. Figure 3), and PPV of 0.57, sensitivity of 0.95, and specificity of 0.67 for the scenario when definite, probable, and doubtful asthma cases were combined and considered relevant for the study (cf. Figure 4). Our experiments with subsets of information available in the medical record (codes, medications, free text) indicate that, overall, best classification performances are obtained with an algorithm that uses all information in the medical record.

Interestingly, none of the 7 rules in the case detection algorithm generated for definite asthma contains an ICPC code for asthma, i.e., R96, or any R03drug (cf. Table 2). The presence of the

terms ‘flixotide’, ‘ventolin’, and ‘pulmicort’, which are all R03drugs, suggests that the specialists’ letters do not (or not very often) contain ATC drug codes. The RIPPER algorithm was able to pick up both the terms used to indicate the specialty of pediatrics or a pediatrician in the IPCI database, i.e., ‘kindergeneeskunde’ and the IPCI database code ‘20’. Similarly, the algorithm also picked up both the terms used for the specialty of pulmonary diseases or a pulmonologist, i.e., ‘longziekten’ and the IPCI database code ‘15’. For probable and doubtful asthma cases, the algorithm picked up both the ICPC asthma code R96 and R03drug (cf. Table 3). The algorithm was also able to pick up specific drug names such as ‘flixotide’, ‘ventolin’, and ‘pulmicort’ and abbreviations such as ‘inh’ for ‘inhaler’ and ‘vag’ for ‘vesiculair ademgeruis’ (vesicular breath sound) used within the IPCI database. A comparison with the broad query shows that the automated case definitions contain more specific keywords (and combinations) used within the database. This suggests that rules with database specific keywords are complicated to construct manually for use in the broad query.

There were some study limitations. The RIPPER algorithm used a training set of positive and negative examples of asthma cases from the IPCI database. The generated case detection algorithm is therefore specific to the IPCI database and it may not be applicable to other databases to detect asthma cases. A new training set is required to generate an automated case detection algorithm for a new EHR database. The automated case detection algorithm is applicable within the results of the broad query. The automated case detection algorithm will also miss any asthma case initially missed by the broad query. However, such asthma cases can potentially be identified by applying the automated case detection algorithm onto the complete EHR database, although we do not know how well this would work.

Usually the only way to extract or identify cases from the electronic health record databases is using codes such as ICPC or ICD-9 because the free-text narratives cannot be easily analyzed. Recently, Flynn et al. [29] used free-text clinical reports to develop an algorithm using manual rules to identify ischaemic stroke and intracerebral haemorrhage. The approach we used in this study to generate a case detection algorithm to identify asthma patients has a number of advantages. Our approach not only used the structured information, as is usually done, but also took advantage of the free-text narratives present in the EHR database. Another advantage relates to patient confidentiality, which is a matter of concern when dealing with free-text in electronic health records. In our approach, once a model has been generated, cases can automatically be identified without need to anonymize data. We also demonstrated how sensitivity and specificity of the algorithms can be tuned to best meet the requirements of our study. An automatic case detection algorithm with high specificity can reduce the workload of manual annotation, by removing non-relevant records. Another advantage of automated case detection algorithm is that they can allow for more uniform and consistent annotations as compared to several manual annotators. Although the case detection algorithm for asthma discussed here is specific to the IPCI database, the approach used to generate the algorithm can be used in different databases.

In databases such as IPCI, manual review of all results of the broad query is currently mandatory in order to identify asthma cases. Using the automated algorithm described here, it is now feasible to automatically identify definite, probable, and doubtful asthma patients with

acceptable performance, using both free-text narratives and coded information when available, allowing large scale epidemiology studies.

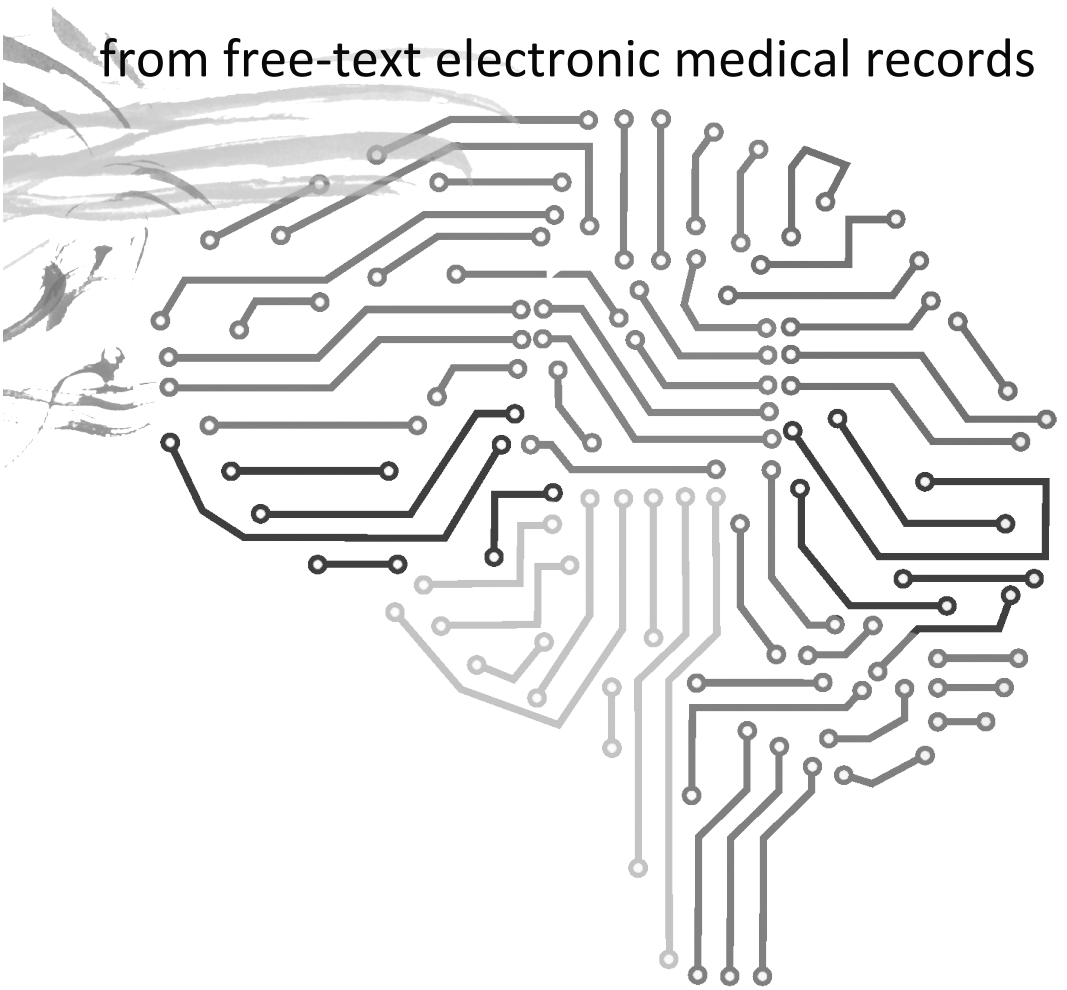
REFERENCES

1. Bateman ED, Hurd SS, Barnes PJ, Bousquet J, Drazen JM, FitzGerald M, Gibson P, Ohta K, O'Byrne P, Pedersen SE, Pizzichini E, Sullivan SD, Wenzel SE, Zar HJ: **Global strategy for asthma management and prevention: GINA executive summary.** *Eur Respir J Off J Eur Soc Clin Respir Physiol* 2008, **31**:143–78.
2. Wallerstedt SM, Brunlöf G, Sundström A, Eriksson AL: **Montelukast and psychiatric disorders in children.** *Pharmacoepidemiol Drug Saf* 2009, **18**:858–864.
3. Cates CJ, Cates MJ: **Regular treatment with formoterol for chronic asthma: serious adverse events.** *Cochrane Database Syst Rev* 2012, **4**:CD006923.
4. Cates CJ, Cates MJ: **Regular treatment with salmeterol for chronic asthma: serious adverse events.** *Cochrane Database Syst Rev* 2008:CD006363.
5. Allen DB: **Effects of Inhaled Steroids on Growth, Bone Metabolism, and Adrenal Function**. *Adv Pediatr* 2006, **53**:101–110.
6. Kelly HW, Van Natta ML, Covar RA, Tonascia J, Green RP, Strunk RC: **Effect of long-term corticosteroid use on bone mineral density in children: a prospective longitudinal assessment in the childhood Asthma Management Program (CAMP) study.** *Pediatrics* 2008, **122**:e53-61.
7. Incecik F, Onlen Y, Sangun O, Akoglu S: **Probable montelukast-induced hepatotoxicity in a pediatric patient: case report.** *Ann Saudi Med* 2007, **27**:462–463.
8. Sass DA, Chopra KB, Wu T: **A case of montelukast-induced hepatotoxicity.** *Am J Gastroenterol* 2003, **98**:704–705.
9. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW: **Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting**. *Pharmacoepidemiol Drug Saf* 2010, **19**:1211–1215.
10. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR: **Temporal pattern discovery in longitudinal electronic patient records**. *Data Min Knowl Discov* 2009, **20**:361–387.
11. Boockvar KS, Livote EE, Goldstein N, Nebeker JR, Siu A, Fried T: **Electronic health records and adverse drug events after patient transfer.** *Qual Saf Health Care* 2010, **19**:e16.
12. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW: **Electronic health record-based cardiac risk assessment and identification of unmet preventive needs.** *Med Care* 2009, **47**:418–424.
13. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM: **Electronic medical records for discovery research in rheumatoid arthritis.** *Arthritis Care Res (Hoboken)* 2010, **62**:1120–1127.
14. Allen-Dicker J, Klompas M: **Comparison of electronic laboratory reports, administrative claims, and electronic health record data for acute viral hepatitis surveillance.** *J Public Health Manag Pract* 2012, **18**:209–214.
15. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA: **Automating classification of free-text electronic health records for epidemiological studies.** *Pharmacoepidemiol Drug Saf* 2012, **21**:651–658.
16. Vlug A, van der Lei J, Mosseveld B, van Wijk M, van der Linden P, MC S, van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project**. *Methods Inf Med* 1999, **38**:339–344.
17. Lamberts H, Wood M: **ICPC: International Classification of Primary Care**. *Scand J Prim Heal Care* 1987:204.

18. **Snowball stemmers**. :http://snowball.tartarus.org/.
19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: **A simple algorithm for identifying negated findings and diseases in discharge summaries.** *J Biomed Inform* 2001, **34**:301–310.
20. Setiono R, Liu H: **Chi2: feature selection and discretization of numeric attributes**. In *Proc 7th IEEE Int Conf Tools with Artif Intell*. IEEE Comput. Soc. Press; 1995:388–391.
21. Kiritchenko S, Matwin S, Nock R, Famili AF: **Learning and evaluation in the presence of class hierarchies: Application to text categorization**. *Adv Artif Intell* 2006, **4013**:395–406. [*Lecture Notes in Computer Science*]
22. Costa EP, Lorena AC, Carvalho ACPLF, Freitas AA, Holden N: **Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees**. In *Adv Bioinforma Comput Biol*. *Volume 4643*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007:126–137.
23. Metz J, Freitas AA, Monard MC, Cherman EA: **A study on the selection of local training sets for hierarchical classification tasks**. In *Brazilian Natl Meet Artif Intell*. Natal, RN, Brasil: Sociedade Brasileira da Computa - SBC; 2011:572–583.
24. Chawla N V: **Data Mining for Imbalanced Datasets: An Overview**. In *Data Min Knowl Discov Handb*. Edited by Maimon O, Rokach L. Boston, MA: Springer US; 2010:875–886.
25. Cohen WW: **Fast Effective Rule Induction**. In *Proc Twelfth Int Conf Mach Learn*. Edited by Prieditis A, Russell S. Tahoe City, California, USA: Morgan Kaufmann; 1995:115–123.
26. Geurts P, Irrthum A, Wehenkel L: **Supervised learning with decision tree-based methods in computational and systems biology**. *Mol Biosyst* 2009, **5**:1593–1605.
27. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software**. *ACM SIGKDD Explor Newsl* 2009, **11**:10.
28. Adler W, Brenning A, Potapov S, Schmid M, Lausen B: **Ensemble classification of paired data**. *Comput Stat Data Anal* 2011, **55**:1933–1941.
29. Flynn RW V, Macdonald TM, Schembri N, Murray GD, Doney ASF: **Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes.** *Pharmacoepidemiol Drug Saf* 2010, **19**:843–7.

Chapter 6

Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records



ABSTRACT

Background

Distinguishing cases from non-cases in free-text electronic medical records is an important initial step in observational epidemiological studies, but manual record validation is time-consuming and cumbersome. We compared different approaches to develop an automatic case identification system with high sensitivity to assist manual annotators.

Methods

We used four different machine-learning algorithms to build case identification systems for two data sets, one comprising hepatobiliary disease patients, the other acute renal failure patients. To improve the sensitivity of the systems, we varied the imbalance ratio between positive cases and negative cases using under- and over-sampling techniques, and applied cost-sensitive learning with various misclassification costs.

Results

For the hepatobiliary data set, we obtained a high sensitivity of 0.95 (on a par with manual annotators, as compared to 0.91 for a baseline classifier) with specificity 0.56. For the acute renal failure data set, sensitivity increased from 0.69 to 0.89, with specificity 0.59. Performance differences between the various machine-learning algorithms were not large. Classifiers performed best when trained on data sets with imbalance ratio below 10.

Conclusions

We were able to achieve high sensitivity with moderate specificity for automatic case identification on two data sets of electronic medical records. Such a high-sensitive case identification system can be used as a pre-filter to significantly reduce the burden of manual record validation.

Background

Electronic medical records (EMRs) are nowadays not only used for supporting the care process, but are often reused in observational epidemiological studies, e.g., to investigate the association between drugs and possible adverse events [1-3]. An important initial step in these studies is case identification, i.e., the identification of patients who have the event of interest. Case identification is particularly challenging when using EMRs because data in the EMRs are not collected for this purpose [4]. Ideally, case identification is done on data that have been coded explicitly and correctly with a structured terminology such as the International Classification of Diseases version 9 (ICD-9). However, coding is often not available. For example, in the Integrated Primary Care Information (IPCI) database [5] used in this study, almost 60% of the record lines comprise only narratives and no coded information. The non-coded part contains essential information, such as patient-reported symptoms, signs, or summaries of specialists' letters in narrative form. This information may be critical for identification of the events. The use of non-coded data (along with the coded data) in medical records has been shown to significantly improve the identification of cases [6]. However, the most commonly used method for case identification is using coded data only [7-11]. The current workflow of epidemiological case identification typically consists of two steps: 1) issuing a broad query based on the case definition to select all potential cases from the database, and 2) manually reviewing the patient data returned by the query to distinguish true positive cases from true negative cases. Manual review of the patient data is an expensive and time-consuming task, which is becoming prohibitive with the increasing size of EMR databases. Based on our recorded data, on average about 30 patients are reviewed per hour by a trained annotator. For a data set of 20,000 patients, which is an average-sized data set in our studies, almost 650 hours (~90 days) will be required. To make case identification more efficient, manual procedures should be replaced by automated procedures as much as possible. Machine learning techniques can be employed to automatically learn case definitions from an example set of free-text EMRs. It is crucial that an automatic case identification system does not miss many positive cases, i.e., it should have a high sensitivity. This is particularly important in incidence rate studies where the goal is to find the number of new cases in a population in a given time period. Any false-positive cases returned by the system would have to be filtered out manually, and thus the classifier should also have a good specificity, effectively reducing the workload considerably as compared to a completely manual approach.

There is a substantial amount of literature on identifying and extracting information from EMRs [12]. Machine-learning methods have been used for different classifications tasks based on electronic medical records such as identification of patients with various conditions [6,13-17], automatic coding [7,18,19], identifying candidates in need of therapy [20], identifying clinical entries of interest [21], and identifying smoking status [22,23]. Schuemie et al. [24] compared several machine-learning methods for identifying patients with liver disorder from free-text medical records. These methods are usually not optimized for sensitivity but for accuracy. The topic of automatic case identification with high sensitivity has not yet been addressed.

Typically, the proportion of positive and negative cases in a data set is not equal (usually there are many more negative cases than positive cases). This imbalance affects the learning process [25]. We use two approaches to deal with the imbalance problem: sampling methods and cost-sensitive learning. Sampling methods change the number of positive or negative cases in the data set to balance their proportions improving classifiers accuracy. This is achieved by removing the majority class examples, known as under-sampling, or by adding to the minority class examples, known as over-sampling. Both under and over-sampling methods have their drawbacks as well. Under-sampling can remove some important examples from the dataset whereas over-sampling can lead to overfitting [26]. Over- and under-sampling methods, with several variations, have been successfully used to deal with imbalanced data sets [27-32]. It has also been shown that a simple random sampling method can perform equally well as some of the more sophisticated methods [33]. We propose a modified random sampling strategy to boost sensitivity. Cost-sensitive learning tackles the imbalance problem by changing the misclassification costs [34-37]. Cost-sensitive learning is shown to perform better than sampling methods in some application domains [38-40].

In this article, we focus on improving the sensitivity of machine-learning methods for case identification in epidemiological studies. We do this by dealing with the balance of positive and negative cases in the data set, which in our case consists of all potential patients returned by the broad query. A highly sensitive classifier with acceptable specificity can be used as a pre-filter in the second step of the epidemiology case identification workflow to distinguish positive cases and negative cases. The experiments are done on two epidemiological data sets using four machine-learning algorithms.

Methods

Data sets

Data used in this study were taken from the IPCI database [5]. The IPCI database is a longitudinal collection of EMRs from Dutch general practitioners containing medical notes (symptoms, physical examination, assessments, and diagnoses), prescriptions and indications for therapy, referrals, hospitalizations, and laboratory results of more than 1 million patients throughout the Netherlands. A patient record consists of one or more entries, where each entry pertains to a patient visit or a letter from a specialist.

We used two data sets, one with hepatobiliary disease patients and one with acute renal failure patients. These data sets are very different from each other and are taken from real-life drug-safety studies in which it is important to investigate the incidence and prevalence of the outcomes in the general population. This type of studies serves as a good example for building highly sensitive automatic case identification algorithm because they require that all the cases in the population are identified. To construct the data sets, first a broad query was issued to the IPCI database. The aim of the query was to retrieve all potential cases according to the case definition. The query included any words, misspellings, or part of the words relevant to the case

definition. The sensitivity of the broad query is very high but its specificity is usually low, and therefore many of the cases retrieved by the query are likely to be negative cases.

To train the machine-learning algorithms, a random sample of the entries returned by the broad query was selected. The size of the random sample may depend on the complexity of the case definition and the disease occurrence. Our experience suggests that the size of the random sample should be a minimum of 1,000 entries to get good performance. All patients pertaining to the randomly selected entries were manually labeled as either positive or negative cases. Because the broad query might have returned an entry with circumstantial evidence but have missed the entry with the actual evidence (e.g., because of textual variation in keywords), the entire medical record (all entries) of the patients in the random sample was considered to decide on a label, not only the entry returned by the broad query. A patient was labeled as a positive case if evidence for the event was found in any of the patient's entries. The patient was labeled as a negative case if there was no proof of the event in any of the patient's entries.

Each random sample was manually labeled by one medical doctor. These labels are used as a gold standard. To verify the quality of the labels and to calculate inter-observer agreement, another medical doctor then labeled a small random set (n=100) from each random sample. We used Cohen's Kappa to calculate the agreement between both annotators [41].

Hepatobiliary disease was defined as either gallstones (with or without surgery), cholecystitis, hepatotoxicity, or general hepatological cases such as hepatitis or liver cirrhosis. The broad query retrieved 53,385 entries, of which 1,000 were randomly selected for manual labeling. These 1,000 entries pertained to 973 unique patients, of whom 656 were labeled as positive cases of hepatobiliary disease and 317 were labeled as negative cases.

Acute renal failure was defined as a diagnosis of (sub)acute kidney failure/injury/insufficiency by a specialist and hospitalization, or renal replacement therapy followed by acute onset of sepsis, operation, shock, reanimation, tumorlysissyndrome, or rhabdomyolysis. The broad query for acute renal failure patients retrieved 9,986 entries, pertaining to 3,988 patients who were all manually labeled. Only 237 patients were labeled as positive cases of acute renal failure and 3,751 patients were labeled as negative cases. Of these latter, many had chronic renal failure.

The labeled set included one entry per patient. For positive cases, we selected the entry with the evidence or, if multiple such entries were available, one was randomly chosen. For negative cases, we randomly selected an entry. The selected entries will be called 'seen entries' from here onwards.

Preprocessing

Since a medical record may contain differential diagnosis information, it is important to distinguish between positive statements made by the physician, and negations and perhaps speculations. In order to remove negated and speculative assertions we use an assertion filter, similar to others [42]. We identify three sets of keywords:

- Speculation keywords: Words indicating a speculation by the physician (e.g. 'might', 'probable', or 'suspected')

- Negation keywords: Words indicating a negation (e.g. 'no', 'not', or 'without')
- Alternatives keywords: Words indicating potential alternatives (e.g. 'versus', or 'or')

Note that the medical records and these keywords are in Dutch. Any words appearing between negation or speculation keywords and the end of a sentence (demarked by a punctuation mark) were removed from the record. Similarly, all sentences containing an alternatives keyword were completely removed. The remaining text was converted to lower case and split into individual words.

After the removal of negation, speculation, and alternative assertions, all remaining individual words in an entry were treated as features (bag-of-words representation). The advantage of using the assertion filter and bag-of-words feature representation on Dutch EMRs is presented in [24]. Since the total number of features was still very high even after preprocessing, which makes machine learning computationally expensive and may also hamper the predictive accuracy of the classifier, we performed chi-square feature selection [43]. For each feature, we compared the feature distribution of the cases and non-cases by a chi-square test. If the test was significant, the feature was selected for further processing. A p-value of less than 0.05 was used as feature selection threshold. Feature selection was done as a preprocessing step in each of the cross-validation training folds of the data sets.

Set expansion

Adding more cases (i.e. patients) in the data set is expensive because they have to be first manually validated and labeled. We used 'set expansion' as an alternative approach to expand the training and test set. Each labeled set consisted of positive and negative cases, one (seen) entry per case. The fact that each case typically has multiple entries, allowed us to expand the labeled sets. For a negative case, the annotator has extensively reviewed all of the entries in the patient record and found no convincing positive evidence. Although only one random entry (seen entry) was selected for a negative case, we can however use all other entries as additional negative examples for the machine learning because none of them contained any convincing positive evidence. We call these additional negative examples the 'implicit entries'. For a positive case, the annotator selected an entry containing convincing positive evidence (seen entry). For all other entries of a positive case, it is uncertain whether these entries also contain convincing positive evidence. These entries therefore cannot be used as positive examples for the machine learning. We call these uncertain entries of positive cases the 'unseen entries'.

Training and testing

To train and test our classifiers, we used 5-fold cross-validation. Cross validation was done at the patient level (subject-level cross-validation [44]), i.e., the data set was randomly divided in five equally sized subsets of cases. In five cross-validation runs, each time the entries pertaining to four subsets of cases were used as a training set and the entries of the remaining subset were used for testing. For training, we used two sets of entries: a set without set expansion (i.e., with only the seen entries) and a set with set expansion (i.e., with seen and implicit entries). For testing

the classifiers, however, we used all entries of the patients in the test fold. The numbers of seen, implicit, and unseen entries per data set are summarized in Table 1.

Table 1: Total number of subjects and corresponding entries in the hepatobiliary disease and acute renal failure data sets

	Hepatobiliary disease	**Acute renal failure**
Positive cases	656	237
Seen entries	656	237
Unseen entries	61,179	58,022
Negative cases	317	3,751
Seen entries	317	3,751
Implicit entries	27,276	319,204

All entries of the patients in the test fold were used to simulate a real-life situation where we do not know the labels of the entries pertaining to the patients returned by the broad query. We chose not to limit ourselves to the entries returned by the broad query as they may not always contain the entry with evidence (see above), but always included all entries available for each case in the test fold.

We used sensitivity and specificity measures to evaluate the performance of the classifiers. Sensitivity is defined as the true-positive recognition rate: number of true positives / (number of true positives + number of false negatives), whereas specificity is defined as the true-negative recognition rate: number of true negatives / (number of true negatives + number of false positives).

Improving classifiers sensitivity

The imbalance of positive and negative examples in the training set effects the classifiers performance [23]. We used sampling and cost-sensitive learning approaches to improve the sensitivity of our classifiers by dealing with this imbalance.

Sampling

Given an initially imbalanced data set, our proposed random sampling strategy focuses on increasing the proportion of positive case entries in the data set. Because the standard classifiers are biased towards the majority class [45-47], this improvement will potentially help the learning algorithms to generate models that better predict the positive cases, and thus improve sensitivity. In under-sampling, we only removed entries of negative cases regardless of their being in the majority or minority. For the data set with set expansion, under-sampling was done only on the implicit entries (cf. Table 1), varying from 10% under-sampling to 100% (all implicit

entries removed). Thus, each negative case was left with at least one entry (the seen entry). For the data set without set expansion, under-sampling was done on the seen entries, effectively removing negative cases from the data set.

In our random over-sampling approach, we duplicated the entries of positive cases, regardless of their being in the majority or minority. The number of entry duplications was varied between 1 and 10.

Cost-sensitive learning

Cost-sensitive learning methods can be categorized into two categories, direct methods and meta-learning or wrapper methods [34]. In direct cost-sensitive learning, the learning algorithm takes misclassification costs into account. These types of learning algorithms are called cost-sensitive algorithms. In meta-learning, any learning algorithm, including cost-insensitive algorithms, is made cost-sensitive without actually modifying the algorithm.

We chose to use MetaCost [48], a meta-learning approach, in its Weka implementation [49]. Given a learning algorithm and a cost matrix, MetaCost generates multiple bootstrap samples of the training data, each of which is used to train a classifier. The classifiers are then combined through a majority-voting scheme to determine the probability of each example belonging to each class. The original training examples in the data set are then relabeled based on a conditional risk function and the cost matrix [48]. The relabeled training data are then used to create a final classifier.

The cost of misclassification is often not known and there are no standard guidelines available for setting up the cost matrix. Some researchers have used the ratio of positives to negatives as the misclassification cost (20) but this has been questioned by others (21). The values in the cost matrix are also dependent on the base classifier used. Some classifiers require a small misclassification cost while others require a large misclassification cost to achieve the same result. In our experiments, we varied the misclassification costs from 1 to 1000 in 9 steps.

Classifiers

We selected the four top-performing algorithms from a previous study [24], in which many well-known machine-learning algorithms were evaluated for the classification of EMRs in a similar experimental setting.

- C4.5 [50], a well-known decision-tree learner. Weka's implementation of C4.5 (called J48) is used in the experiments.
- Support Vector Machines (SVM) [46], a commonly used algorithm that can handle large data sets. Weka's implementation of libsvm [51] is used in the experiments. Because we had a large number of binary features, we used a linear kernel [52] and the soft margin parameter *c* was set to 4.

- RIPPER [53], a decision-rule learner. RIPPER induces an ordered set of rules by combining covering with a reduced error pruning strategy. Weka's implementation of RIPPER (called JRip) is used in the experiments.
- MyC, a locally developed decision-tree learner. MyC builds a tree by iteratively splitting the data based on the chi-square test, similar to the ID3 algorithm [54]. MyC is simple and very fast.

We did an error analysis to understand why some of the positive cases were not identified by the classifiers. Errors were divided in the following four categories: evidence keywords not picked up by the algorithm, evidence keyword picked up by the algorithm but removed from the patient entry by the negation/speculation filter, different spelling variations of the evidence keywords in the learned model and in the evidence entry, and patient wrongly labeled as a positive case by the annotator.

Results

There was a good to excellent agreement between the two annotators (kappa scores of 0.74 (95% CI 0.59-0.89) and 0.90 (95% CI 0.83-0.97) for the hepatobiliary and acute renal failure data sets, respectively). The chi-square feature selection decreased the number of features in both data sets by about a factor of 10, without affecting the performance of the classifiers but greatly reducing their training time. For example, RIPPER using MetaCost took about five days to build one classifier for the acute renal failure set, which after feature selection took less than one day.

Table 2 shows the sensitivity and specificity results of all four classifiers trained on the hepatobiliary and the acute renal failure data sets, with and without set expansion.

Table 2: Sensitivity and specificity results of various classifiers trained on the hepatobiliary and the acute renal failure data sets, with and without set expansion

	Set	Imbalance	SVM		C4.5		MyC		RIPPER	
Data set	expansion	ratio	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Hepatobiliary	No	0.5	0.99	0.03	0.99	0.03	0.99	0.07	0.99	0.04
	Yes	42	0.89	0.77	0.90	0.79	0.92	0.69	0.91	0.71
Acute renal failure	No	16	0.62	0.92	0.69	0.88	0.69	0.90	0.71	0.89
	Yes	1363	0.39	0.98	-	-	0.45	0.99	0.41	0.98

C4.5 could not generate a classifier for our largest data set, acute renal failure with set expansion, because the memory requirement of this algorithm proved prohibitive.

The decision-tree and decision-rule learners performed slightly better than the SVM. The imbalance ratios (number of negative examples divided by number of positive examples) varies greatly for the baseline classifiers. The specificity of the classifiers trained on the hepatobiliary data without set expansion was very low. For our sampling and cost-sensitive experiments, we therefore focused on changing the imbalance ratio in the data with set expansion. The acute renal failure data with set expansion was very imbalanced, which resulted in classifiers with

relatively low sensitivity. We therefore focused on changing the imbalance ratio in the data without set expansion.

Tables 3, 4, 5 and 6 show the results for changing the proportions of positive and negative cases in both data sets by under-sampling and over-sampling, respectively.

Table 3: Sensitivity and specificity of various classifiers trained on the hepatobiliary data set for difference percentages of under-sampling

Under-sampling	SVM		MyC		RIPPER		C4.5		Imbalance
(%)	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	ratio
0	0.89	0.77	0.92	0.68	0.91	0.71	0.90	0.79	42
10	0.89	0.76	0.93	0.65	0.91	0.75	0.90	0.80	38
20	0.89	0.75	0.93	0.63	0.91	0.73	0.91	0.79	34
30	0.89	0.76	0.94	0.61	0.93	0.72	0.90	0.78	30
40	0.89	0.73	0.93	0.60	0.92	0.69	0.91	0.77	25
50	0.90	0.70	0.93	0.58	0.92	0.71	0.91	0.76	21
60	0.90	0.71	0.94	0.56	0.92	0.72	0.92	0.73	17
70	0.91	0.67	0.95	0.55	0.91	0.72	0.92	0.70	13
80	0.92	0.64	0.94	0.49	0.92	0.73	0.92	0.68	9
90	0.94	0.52	0.91	0.60	0.93	0.67	0.93	0.59	5
100	0.99	0.12	0.99	0.07	0.99	0.03	0.99	0.14	0.5

Table 4: Sensitivity and specificity of various classifiers trained on the acute renal failure data set for difference percentages of under-sampling

Under-sampling	SVM		MyC		RIPPER		C4.5		Imbalance
(%)	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	ratio
0	0.62	0.92	0.69	0.90	0.71	0.89	0.69	0.88	16
10	0.64	0.90	0.74	0.89	0.75	0.89	0.69	0.87	14
20	0.64	0.89	0.75	0.83	0.75	0.88	0.74	0.86	13
30	0.66	0.88	0.76	0.82	0.76	0.88	0.75	0.85	11
40	0.70	0.85	0.75	0.87	0.74	0.88	0.75	0.85	9
50	0.74	0.81	0.76	0.80	0.77	0.76	0.76	0.82	8
60	0.82	0.72	0.77	0.81	0.84	0.68	0.83	0.82	6
70	0.83	0.67	0.83	0.70	0.83	0.61	0.86	0.77	5
80	0.86	0.56	0.89	0.49	0.90	0.44	0.90	0.45	3
90	0.92	0.41	0.90	0.43	0.89	0.43	0.92	0.39	2

Table 5: Sensitivity and specificity of various classifiers trained on the hepatobiliary data set for difference percentages of over-sampling

Over-sampling	SVM		MyC		RIPPER		C4.5		Imbalance
(%)	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	ratio
0	0.89	0.77	0.92	0.68	0.91	0.71	0.90	0.79	42
100	0.90	0.72	0.96	0.52	0.94	0.64	0.93	0.73	21
200	0.90	0.70	0.96	0.47	0.96	0.56	0.94	0.67	14
300	0.91	0.70	0.97	0.44	0.96	0.54	0.95	0.65	11
400	0.91	0.71	0.98	0.45	0.97	0.50	0.95	0.63	8
500	0.92	0.69	0.98	0.43	0.97	0.48	0.95	0.62	7
600	0.92	0.68	0.97	0.35	0.96	0.47	0.95	0.61	6
700	0.92	0.67	0.98	0.34	0.97	0.47	0.95	0.60	5
800	0.92	0.65	0.97	0.34	0.97	0.47	0.95	0.61	5
900	0.93	0.65	0.97	0.34	0.97	0.45	0.95	0.59	4
1000	0.93	0.64	0.97	0.35	0.96	0.44	0.95	0.59	4

Table 6: Sensitivity and specificity of various classifiers trained on the acute renal failure data set for difference percentages of over-sampling

Over-sampling	SVM		MyC		RIPPER		C4.5		Imbalance
(%)	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	ratio
0	0.62	0.92	0.69	0.90	0.75	0.89	0.69	0.88	16
100	0.66	0.86	0.78	0.80	0.81	0.76	0.74	0.75	8
200	0.71	0.81	0.84	0.71	0.84	0.65	0.77	0.67	5
300	0.74	0.77	0.89	0.59	0.88	0.65	0.80	0.65	4
400	0.76	0.73	0.89	0.51	0.86	0.64	0.81	0.61	3
500	0.77	0.69	0.89	0.48	0.84	0.64	0.82	0.60	3
600	0.78	0.66	0.91	0.48	0.89	0.59	0.82	0.60	2
700	0.82	0.60	0.92	0.43	0.89	0.54	0.82	0.60	2
800	0.82	0.57	0.94	0.37	0.86	0.60	0.82	0.61	2
900	0.83	0.55	0.93	0.36	0.89	0.53	0.83	0.61	2
1000	0.84	0.54	0.95	0.36	0.88	0.54	0.83	0.61	1

All algorithms showed consistent behavior during the under-sampling experiments. The sensitivity increased and specificity decreased as we decrease the number of negative case entries from the data set.

Almost a similar pattern is observed during the over-sampling experiments where sensitivity gradually increased and specificity decreased as we increase the number of positive case entries in the data set. MyC showed slightly more improvement in the sensitivity as compared to other algorithms but then also lower specificity.

The results for cost-sensitive learning with MetaCost using varying misclassification costs are shown in Tables 7 and 8.

Table 7: Sensitivity and specificity of various classifiers trained on the hepatobiliary data set for difference cost values of cost-sensitive learning

	SVM		MyC		RIPPER		C4.5	
Cost	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**
1	0.86	0.78	0.90	0.68	0.93	0.67	0.89	0.71
10	0.87	0.78	0.95	0.54	0.93	0.68	0.92	0.69
25	0.87	0.79	0.96	0.47	0.93	0.67	0.92	0.69
50	0.87	0.79	0.96	0.47	0.93	0.67	0.91	0.66
100	0.87	0.79	0.96	0.47	0.93	0.67	0.92	0.66
200	0.87	0.79	0.96	0.47	0.93	0.67	0.92	0.66
400	0.87	0.79	1.00	0.09	0.97	0.24	0.99	0.12
800	0.87	0.79	1.00	0.00	1.00	0.00	1.00	0.00
1000	0.87	0.79	1.00	0.00	1.00	0.00	1.00	0.00

Table 8: Sensitivity and specificity of various classifiers trained on the acute renal failure data set for difference cost values of cost-sensitive learning

	SVM		MyC		RIPPER		C4.5	
Cost	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**
1	0.59	0.92	0.74	0.85	0.78	0.80	0.67	0.73
10	0.59	0.92	0.81	0.63	0.78	0.80	0.73	0.69
25	0.59	0.92	0.81	0.63	0.78	0.80	0.76	0.64
50	0.59	0.92	0.89	0.35	0.78	0.80	0.78	0.60
100	0.59	0.92	1.00	0.00	0.78	0.80	0.97	0.11
200	0.59	0.92	1.00	0.00	1.00	0.00	1.00	0.00
400	0.59	0.92	1.00	0.00	1.00	0.00	1.00	0.00
800	0.59	0.92	1.00	0.00	1.00	0.00	1.00	0.00
1000	0.59	0.92	1.00	0.00	1.00	0.00	1.00	0.00

Classifiers do not seem to be very sensitive to the misclassification cost so performance variations were observed at relatively high cost values.

As an example of the sensitivity that can be achieved with the sampling methods and cost-sensitive learning while maintaining a reasonable specificity, Table 9 shows the performance of

the classifiers with the highest sensitivity and a specificity of at least 0.5. Our results (cf. Tables 3, 4, 5, 6, 7 and 8) show that classifiers with high specificity than 0.5 are feasible but at the expense of a lower sensitivity.

Table 9: Performance of the classifiers with the highest sensitivity and a specificity of at least 0.5 on the hepatobiliary disease and acute renal failure data sets

Data set	**Algorithm**	**Baseline**		**Under-sampling**		**Over-sampling**		**Cost-sensitive**	
		Sens.	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**
Hepatobiliary disease	SVM	0.89	0.77	**0.94**	**0.52**	0.93	0.65	0.87	0.79
	MyC	0.92	0.68	**0.95**	**0.56**	0.94	0.54	0.95	0.54
	C4.5	0.90	0.79	0.93	0.59	**0.94**	**0.56**	0.92	0.66
	RIPPER	0.90	0.71	0.93	0.72	**0.94**	**0.51**	0.93	0.67
Acute renal failure	SVM	0.62	0.92	**0.86**	**0.56**	0.84	0.54	0.59	0.92
	MyC	0.69	0.90	0.83	0.70	**0.89**	**0.51**	0.81	0.63
	C4.5	0.69	0.88	**0.86**	**0.77**	0.83	0.61	0.78	0.60
	RIPPER	0.71	0.89	0.84	0.68	**0.89**	**0.59**	0.78	0.80

Table 10: Error analysis of the false negatives by the MyC classifier trained on the hepatobiliary disease data set with 70% under-sampling

Type of error	**N (%)**
Evidence not in the model	13 (38)
Evidence removed by negation/speculation filter	12 (35)
Spelling variations	5 (15)
Labeling error	4 (12)

The performance of sampling methods and cost-sensitive learning is compared to the baseline models of both data sets.

To get an estimate of the sensitivity and specificity of manual case identification, we compared the labels of the second annotator with the gold standard labels of annotator 1. For the hepatobiliary set, sensitivity was 0.94 and specificity was 0.83, for the acute renal failure set sensitivity was 0.96 and specificity was 0.94. Our experiments (cf. Tables 3, 4, 5, 6, 7 and 8) showed that similar sensitivity performance (or even better sensitivity for the hepatobiliary set,

depending on how much specificity can be compromised in a study) could be achieved using automatic classification.

We did an error analysis of the positive cases missed by the MyC algorithm using 70% under-sampling method (sensitivity 0.95) on the hepatobiliary disease data set. About 38% of the missed positive cases were due to the evidence keywords in the entry (e.g., leverfibrose, hepatomegalie, cholestase) not being picked up by the learning algorithm. For about a third of the missed cases, the negation/speculation filter had erroneously removed the evidence in the entry. For example, in the following entry: "Ron [O] ECHO BB: cholelithiasis, schrompelnier li? X-BOZ: matig coprostase", the evidence "cholelithiasis" was removed by the speculation filter because the sentence ended with a question mark. Spelling variations caused about 15% of the errors (e.g., "levercirrhose" instead of "levercirrose" ("liver cirrhosis"), and 12% of the missed cases turned out to be labeling errors. For example, in the following labeled entry: "Waarschijnlijk steatosis hepatitis bij status na cholecystectomie" the GP has mentioned only a probability of the disease ("waarschijnlijk", meaning "probable"), but the patient was labeled as a positive case.

Discussion

In this paper, we demonstrated that dealing with the proportions of positive and negative cases entries in the data sets could increase the sensitivity of machine learning methods for automated case identification. We used sampling and cost-sensitive methods on two very different data sets and with four different machine-learning algorithms.

The under-sampling and over-sampling methods performed consistently well and resulted in higher sensitivity on both data sets. Although there was no clear winner between under-sampling and over-sampling methods, under-sampling performed slightly better. For the hepatobiliary set, the best sensitivity-specificity score (by selecting the highest value of sensitivity at a specificity larger than 0.5) using over-sampling was 0.94 sensitivity and 0.56 specificity with C4.5, the best score using under-sampling was 0.95 sensitivity and 0.56 specificity with MyC, and the best score using cost sensitive learning was 0.95 sensitivity and 0.54 specificity using MyC (cf. Table 9). For the acute renal failure set, the best sensitivity-specificity score using over-sampling was 0.89 sensitivity and 0.59 specificity using RIPPER, the best score using under-sampling was 0.86 sensitivity and 0.77 specificity using C4.5, and the best score using cost-sensitive learning was 0.81 sensitivity and 0.63 specificity using MyC. Overall, C4.5 and MyC appeared to perform best.

The sampling experiments demonstrated the effect of imbalance in the data sets. The question of finding an optimal or best class distribution ratio has been studied by several researchers in the past [25,55,56]. Our experiments showed that the classifiers performed better (high sensitivity with not too low specificity) when the imbalance ratio (negative cases to positive cases) was below 10 (cf. Tables 3, 4, 5 and 6). This performance improvement between the ratios was observed in both the data sets despite the fact that they were very different from each other.

Previous studies indicate that cost-sensitive learning usually performs as well as sampling methods if not better [39]. In our experiments, cost-sensitive learning performed about equally well as sampling, but it was difficult to find an optimal cost matrix. Different classifiers treat costs differently and finding an optimal cost value depends on the data set and the classifier used.

Another disadvantage of cost-sensitive learning with MetaCost is the large processing time because of its bootstrapping method. For C4.5, which requires high memory and processing capacity, MetaCost did not generate classifiers for our largest data set because processing time became prohibitive.

The positive effect of set expansion for training on the hepatobiliary disease data set can be seen in Table 2. The results show that set expansion of epidemiological data sets with relatively few negative cases can boost specificity with a modest decrease in sensitivity. For example, specificity for C4.5 increased from 0.03 to 0.79 with sensitivity decreasing from 0.99 to 0.90. On this data set, the set expansion compensated for the relatively few negative examples in the data set without set expansion. The set expansion method added new entries (implicit negative case entries, cf. Table 1) with potentially useful features unlike over-sampling, where existing negative entries in the data set would be duplicated, which could lead to the problem of over-fitting. In the acute renal failure data set, negative examples were already in majority in the training model without set expansion. Set expansion further increased the imbalance, which resulted in decreased sensitivity of below 0.5 for all classifiers.

Overall, the decision tree and rule learning algorithms appear to perform slightly better than the statistical algorithms. One important advantage of tree- and rule-learning algorithms is their ability to generate models that are easily interpretable by humans. Such models can be compared with the case definitions created by human experts.

There were some study limitations. The automatic case identification system was applied on the results of the broad query to distinguish positive cases and negative cases. If cases were missed by the broad query, they will also be missed by the automatic system. In other words, the sensitivity of the automatic case identification system is bound by the sensitivity of the broad query. It would be interesting to apply the automatic system on the actual EMR database and compare it with the broad query. The rate of misspellings has shown to be larger in EMRs than in other type of documents [57] but no attempts were made to handle the misspellings in the case identification system. The end of a sentence was demarked by a punctuation mark, which was not optimal as later confirmed, by the error analysis. Our algorithm to find negated and speculative assertions has been developed for the Dutch language and currently is not as sophisticated and comprehensive as some of the algorithms available for English, e.g., NegEx [42] or ConText [58], and ScopeFinder [59]. To deal with such issues, we need to improve our preprocessing methods. The negation algorithm can be made more informative so it can also detect double negations.

Our strategy by dealing with the imbalance ratio in a data set with and without the set expansion will result in a highly sensitive classifier. An acceptable sensitivity-specificity score will depend on the actual requirement and type of the observational study. We would like to point out that our approach is not specific to the IPCI database or the Dutch EMRs used in this study.

Conclusions

We were able to achieve high sensitivity (on a par with the manual annotator) on both data sets using our proposed sampling and cost-sensitive methods. During a case-identification process in

an epidemiological study, all records returned by the broad query need to be manually validated. An automatic case-identification system with high sensitivity and reasonable specificity can be used as a pre-filter to significantly reduce the workload by reducing the amount of records that needs to be manually validated. The specificity can then be increased during the manual validation process on the reduced set. Using manual validation on the reduced set instead of the set retrieved by the broad query could save weeks of manual work in each epidemiological study.

REFERENCE

1. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW: **Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting.** *Pharmacoepidemiol Drug Saf* 2010, **19**:1211–1215.
2. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR: **Temporal pattern discovery in longitudinal electronic patient records.** *Data Min Knowl Discov* 2009, **20**:361–387.
3. Boockvar KS, Livote EE, Goldstein N, Nebeker JR, Siu A, Fried T: **Electronic health records and adverse drug events after patient transfer.** *Qual Saf Health Care* 2010, **19**:e16.
4. Hurdle JF, Haroldsen SC, Hammer A, Spigle C, Fraser AM, Mineau GP, Courdy SJ: **Identifying clinical/translational research cohorts: Ascertainment via querying an integrated multi-source database.** *J Am Med Inform Assoc* 2012, 1–8.
5. Vlug A, Van der Lei J, Mosseveld B, Van Wijk M, Van der Linden P, MC S, Van Bemmel J: **Postmarketing surveillance based on electronic patient records: the IPCI project.** *Methods Inf Med* 1999, **38**:339–344.
6. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM: **Electronic medical records for discovery research in rheumatoid arthritis.** *Arthritis Care Res (Hoboken)* 2010, **62**:1120–1127.
7. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR: **A systematic literature review of automated clinical coding and classification systems.** *J Am Med Inform Assoc* 2010, **17**:646–651.
8. Chung CP, Murray KT, Stein CM, Hall K, Ray WA: **A computer case definition for sudden cardiac death.** *Pharmacoepidemiol Drug Saf* 2010, **19**:563–572.
9. Cunningham A, Stein CM, Chung CP, Daugherty JR, Smalley WE, Ray WA: **An automated database case definition for serious bleeding related to oral anticoagulant use.** *Pharmacoepidemiol Drug Saf* 2011, **20**:560–566.
10. Singh JA, Holmgren AR, Noorbaloochi S: **Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis.** *Arthritis Rheum* 2004, **51**:952–957.
11. Nicholson A, Tate AR, Koeling R, Cassell JA: **What does validation of cases in electronic record databases mean? The potential contribution of free text.** *Pharmacoepidemiol Drug Saf* 2011, **20**:321–324.
12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF: **Extracting information from textual documents in the electronic health record: a review of recent research.** *Yearb Med Inform* 2008, 128–144.
13. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, Cai T, Goryachev S, Zeng Q, Gallagher PJ, Fava M, Weilburg JB, Churchill SE, Kohane IS, Smoller JW: **Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model.** *Psychol Med* 2012, **42**:41–50.
14. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH: **NLP-based identification of pneumonia cases from free-text radiological reports.** *AMIA Annu Symp Proc* 2008, 172–176.
15. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, Kullo IJ: **Discovering peripheral arterial disease cases from radiology notes using natural language processing division of**

biomedical statistics and informatics, 2 division of cardiovascular diseases. 2010, 722–726.

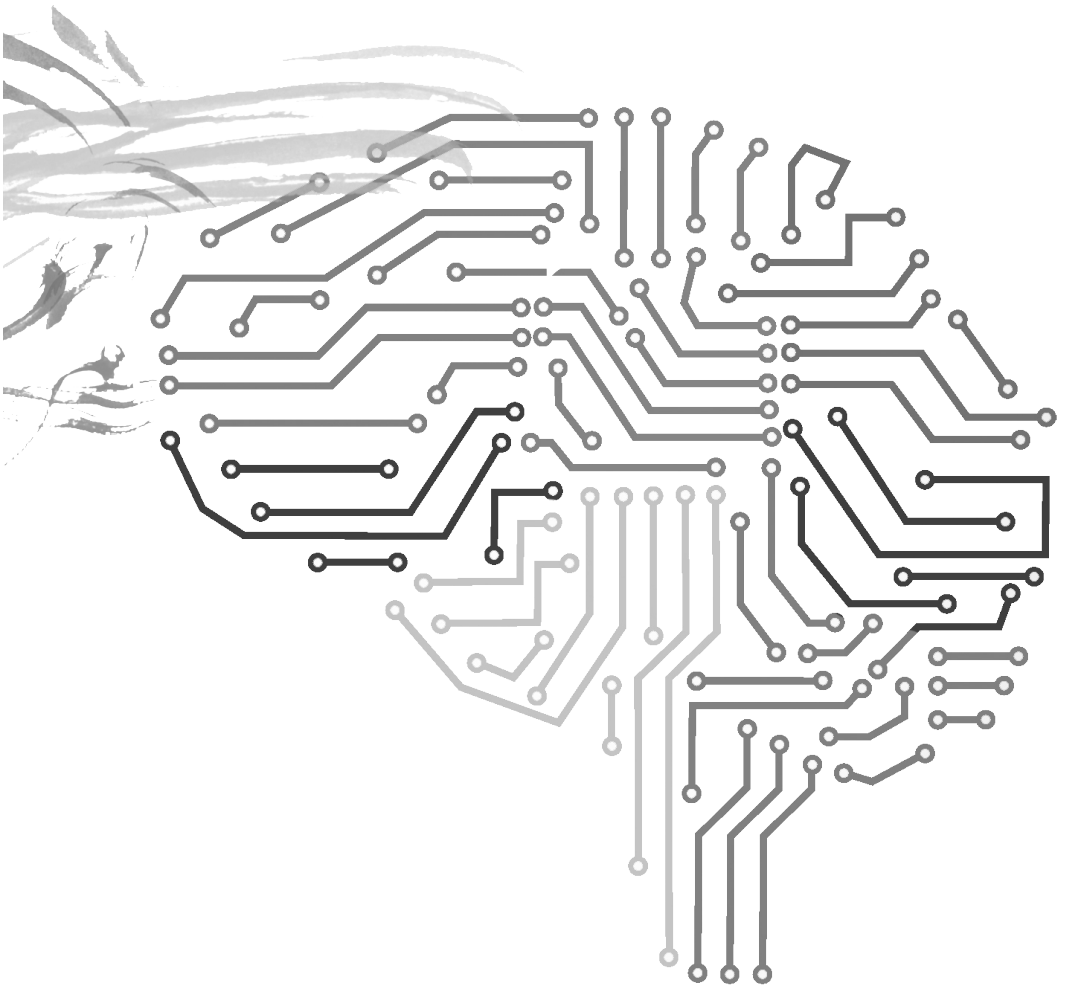
16. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL: **Electronic medical records for clinical research: application to the identification of heart failure.** *Am J Manag Care* 2007, **13**:281–288.
17. Friedlin J, Overhage M, Al-Haddad MA, Waters JA, Aguilar-Saavedra JJR, Kesterson J, Schmidt M: **Comparing methods for identifying pancreatic cancer patients using electronic data sources.** *AMIA Annu Symp Proc* 2010:237–241.
18. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S: **Using electronic patient records to discover disease correlations and stratify patient cohorts.** *PLoS Comput Biol* 2011, **7**:e1002141.
19. Farkas R, Szarvas G: **Automatic construction of rule-based ICD-9-CM coding systems.** *BMC Bioinforma* 2008, **9**(3):S10.
20. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW: **Electronic health record-based cardiac risk assessment and identification of unmet preventive needs.** *Med Care* 2009, **47**:418–424.
21. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H: **Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning.** *PLoS One* 2012, **7**:e30412.
22. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG: **Mayo clinic NLP system for patient smoking status identification.** *J Am Med Inform Assoc* 2008, **15**:25–28.
23. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U: **Identifying smokers with a medical extraction system.** *J Am Med Inform Assoc* 2007, **15**:36–39.
24. Schuemie MJ, Sen E, 't Jong GW, Soest EM, Sturkenboom MC, Kors JA: **Automating classification of free-text electronic health records for epidemiological studies.** *Pharmacoepidemiol Drug Saf* 2012.
25. Garcia EA: **Learning from imbalanced data.** *IEEE Trans Knowl Data Eng* 2009, **21**:1263–1284.
26. Mease D, Wyner AJ: **Boosted classification trees and class probability / quantile estimation.** *J Mach Learn Res* 2007, **8**:409–439.
27. Taft LM, Evans RS, Shyu CR, Egger MJ, Chawla N, Mitchell JA, Thornton SN, Bray B, Varner M: **Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery.** *J Biomed Inform* 2009, **42**:356–364.
28. Van Hulse J, Khoshgoftaar TM, Napolitano A: **An empirical comparison of repetitive undersampling techniques**. In *2009 IEEE International Conference on Information Reuse & Integration.* 2009:29–34.
29. Chawla NV: **Data mining for imbalanced datasets: An overview**. In *Data Mining and Knowledge Discovery Handbook.* Edited by Maimon O, Rokach L. Boston, MA: Springer US; 2010:875–886.
30. Van Hulse J, Khoshgoftaar TM, Napolitano A: **Experimental perspectives on learning from imbalanced data**. In *Proceedings of the 24th international conference on Machine learning - ICML '07.* New York, New York, USA: ACM Press; 2007:935–942.

31. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: synthetic minority over-sampling technique.** *Artif Intell* 2002, **16**:321–357.
32. Drummond C, Holte RC: **C4.5, Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling**. In *Workshop on Learning from Imbalanced Data Sets II (ICML 2003).* 2003:1–8.
33. Japkowicz N: **The class imbalance problem: Significance and strategies**. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI).* 2000:111–117.
34. Ling CX, Sheng VS: *Cost-Sensitive Learning and the Class Imbalance Problem.* In Encyclopedia of Machine Learning: Springer; 2011.
35. Wang T, Qin Z, Zhang S, Zhang C: **Cost-sensitive classification with inadequate labeled data.** *Inf Syst* 2012, **37**:508–516.
36. Japkowicz N, Stephen S: **The class imbalance problem: A systematic study.** *Intell Data Anal* 2002, **6**:429–449.
37. Sun Y, Kamel M, Wong A, Wang Y: **Cost-sensitive boosting for classification of imbalanced data.** *Pattern Recognit* 2007, **40**:3358–3378.
38. Zhou Z, Member S, Liu X: **Training cost-sensitive neural networks with methods addressing the class imbalance problem.** *IEEE Trans Knowl Data Eng* 2006, **18**:63–77.
39. McCarthy K, Zabar B, Weiss G: **Does cost-sensitive learning beat sampling for classifying rare classes?** In *Proceedings of the 1st international workshop on Utility-based data mining - UBDM '05.* New York, New York, USA: ACM Press; 2005:69–77.
40. Liu X, Zhou Z: **The influence of class imbalance on cost-sensitive learning: An empirical study**. In *Sixth International Conference on Data Mining (ICDM'06).* 2006:970–974.
41. Cohen J: **A coefficient of agreement for nominal scales.** *Educ Psychol Meas* 1960, **20**:37–46.
42. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: **A simple algorithm for identifying negated findings and diseases in discharge summaries.** *J Biomed Inform* 2001, **34**:301–310.
43. Setiono R: **Chi2: feature selection and discretization of numeric attributes**. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence.* IEEE Comput. Soc. Press; 1995:388–391.
44. Adler W, Brenning A, Potapov S, Schmid M, Lausen B: **Ensemble classification of paired data.** *Comput Stat Data Anal* 2011, **55**:1933–1941.
45. Sun Y, Kamel M, Wang Y: **Boosting for learning multiple classes with imbalanced class distribution**. In *Sixth International Conference on Data Mining (ICDM'06).* IEEE; 2006:592–602.
46. Akbani R, Kwek S, Japkowicz N: **Applying support vector machines to imbalanced datasets**. In *In Proceedings of the 15th European Conference on Machine Learning (ECML).* 2004:39–50.
47. Chen C, Liaw A, Breiman L: **Using random forest to learn imbalanced data.** *Discovery* 2004, 1–12.
48. Domingos P: **MetaCost: A general method for making classifiers cost-sensitive**. In *Fifth International Conference on Knowledge Discovery and Data Mining.* ACM Press; 1999:155–164.

49. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software**. In *ACM SIGKDD Explorations Newsletter*, Volume 11. 2009:10.
50. Salzberg SL: **C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993.** *Mach Learn* 1994, **16**:235–240.
51. Chang C-C, Lin C-J: **LIBSVM: A library for support vector machines**. In *ACM Transactions on Intelligent Systems and Technology*, Volume 2. 2011:1–27.
52. Hsu C, Chang C, Lin C: **A practical guide to support vector classification.** *Bioinformatics* 2010, **1**:1–16.
53. Cohen WW: **Fast effective rule induction**. In *Proceedings of the Twelfth International Conference on Machine Learning.* Edited by Prieditis A, Morgan Kaufmann RS.; 1995:115–123.
54. Quinlan JR: **Induction of decision trees.** *Mach Learn* 1986, **1**:81–106.
55. Weiss GM, Provost F: **Learning when training data are costly: the effect of class distribution on tree induction.** *J Artif Intell Res* 2003, **19**:315–354.
56. Chan PK, Stolfo SJ: **Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection**. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, Volume 164. 1998:164–168.
57. Ruch P, Baud RH, Geiddbühler A, Lovis C, Rassinoux AM, Rivière A: **Looking back or looking all around: comparing two spell checking strategies for documents edition in an electronic patient record.** *Proc AMIA Symp* 2001, 568–572.
58. Harkema H, Dowling JN, Thornblade T, Chapman WW: **ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports.** *J Biomed Inform* 2009, **42**:839–851.
59. Apostolova E, Tomuro N, Demner-fushman D: **Automatic Extraction of Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics; 2011:283–287.

Chapter 7

Discussion and Conclusion



Electronic health records (EHRs) contain detailed patient information such as demographics, behavior, vital signs, patient-reported symptoms, diagnosis, procedures and treatments, allergies, laboratory data, outpatient and inpatient encounters, and imaging for large patient populations. Although data mining and natural language processing techniques are increasingly being developed and used for automated processing of large amounts of EHRs to answer simple to complex clinical and epidemiological questions, the data-mining tasks remain challenging because of the inherent complexity of the language and diversity of information in the EHRs [1]. Most of the methods and resources developed in the past have focused on English-language EHRs. The work presented in this thesis uses data taken from the Integrated Primary Care Information (IPCI) [2] database, which is a longitudinal collection of EHRs from Dutch general practitioners. The methods and resources developed in this thesis focused particularly on taking advantage of the large unstructured free-text present in the IPCI database. We present our work in the wider framework of knowledge-discovery process, which outlines a standard way of approaching data-mining tasks.

In this chapter, we present an overview and discussion of the findings as described in Chapters 2-6. In the first two sections, we discuss our results in view of the general knowledge-discovery steps of data preparation and data mining. We then discuss the development and use of the Dutch clinical corpus. Finally, we discuss limitations and future work.

Data Preparation

One particular challenge in analyzing free-text EHRs is to distinguish a diagnosis by the physician from the conditions that have been excluded or ruled out. The EHRs often contain family history and past medical problems. Such information should ideally be identified as such. To achieve this, it is important that automated systems do not only identify clinical concepts such as diagnoses but also take into account the context of the identified information. Most work on identifying contextual properties of the clinical concepts has been done on English clinical documents. Recently, the NegEx algorithm [3] has been adapted to detect negations in Swedish [4], French [5], and Spanish [6, 7] clinical text. To our knowledge, no method was yet available for Dutch clinical text. In Chapter 2, we adapted an English language algorithm, ConText [8], which is an extension to NegEx, to the Dutch language in order to identify contextual properties. Such algorithms can be used to identify negated medical concepts and use them as negated features in machine-learning models instead of simply removing them from the training set. We first translated 246 original English language trigger terms into Dutch using Google Translate [9] and then augmented the trigger list manually with additional terms to cover all possible Dutch language specific variations. Compared to the original ConText algorithm, we achieved the same F-score of 0.93 for the negation property when tested on the radiology reports. On discharge letters however, our algorithm, called ContextD, performed better with an F-score of 0.92 as compared to 0.86. The increase in performance can be attributed to our extended list of negation triggers. Although we added an additional module to identify historical properties, our results were lower with an F-score of 0.52 compared to 0.73 for the English version. Our additional temporality module showed a recall of 0.78 as compared to 0.77 for the English version, but it also produced many false positives, which resulted in an overall decrease in the F-score.

Hypothetical property could not be compared since the English version used discharge letters for the evaluation and our discharge letters did not contain any hypothetical concept.

Our results showed that extending the original translated trigger list significantly increases the recall for all properties. Therefore, it could be concluded that while the translated list of English trigger terms provides a good starting point it is important that this is extensively reviewed by a native speaker to cover variations and additional language dependent cases. On the other hand, although our list of Dutch triggers was much longer than the English trigger lists, only a small number of trigger phrases accounted for the majority of the detected terms. This finding is consistent with the findings in other languages [4, 10]. Out of 395 possible Dutch triggers for the negation property, only 23 negation triggers were found in the evaluation set. Most of these 23 negation triggers had an equivalent English trigger term. However, the translation of these terms into Dutch was perhaps not optimal. This could be explained by the fact that we only provided single trigger terms without any context for automated translation and wherever the translator returned more than one alternative, we picked the first one. A context-oriented translation of the English triggers would probably have resulted in better terms. Amongst the three properties, the temporality property appears to be the most difficult one and methods to identify this need to be further developed.

There is no standard way of writing clinical narratives in the EHRs. Often these unstructured clinical narratives are written under time pressure and they contain grammatical errors, standard and non-standard abbreviations, misspellings, ill-formed and incomplete sentences. All automated machine-learning based methods such as case-detection algorithms or rule- and trigger-based algorithms such as ContextD suffer because of such variations. A large number of misspellings or grammar variations greatly increase the number of features in machine-learning tasks when words are used as features, and may reduce the discriminative value of features. In Chapter 3, we considered text normalization as a way to reduce feature dimensions in IPCI-like databases where typographical errors such as spelling mistakes are common. We used three approaches for normalization. In the first approach, we explored the possibilities of using standard terminologies such as SNOMED-CT (Systemized Nomenclature of Medicine – Clinical Terms) and MeSH (Medical Subject Headings) to normalize clinical words in the EHRs. We used Peregrine [11], a concept recognition system, to identify and normalize clinical words to their standard forms in the terminologies. Only a very small percentage of words could be mapped. A likely scenario for low coverage is that the text in IPCI is often noisy and the Peregrine system does not do fuzzy matching, so all words with even a slight variation from the words in the terminology would have been missed. An automatic indexer with fuzzy matching capability would have improved the coverage (although likely at the expense of precision). Another possible reason for low coverage might be that the terminologies we used may just be missing many terms. It can be concluded that for databases like IPCI, where unstructured free-text is often noisy, a terminology-based normalization approach may not be sufficient.

In our second normalization approach, we attempted to automatically identify short-forms (abbreviations or acronyms) and normalize them to their long-form (full form). This topic has been studied before but not so much in the context of clinical text and even less on non-English languages. Normalizing short-form to their long-form is a challenging task in clinical text [12, 13].

Previous approaches mostly involved supervised machine-learning methods such as SVM or CRF where a manually labeled training set is usually available. There is no training set available for this task on the Dutch EHRs and it is expensive and time-consuming to develop a new set. We used heuristics to first identify potential short-forms from the EHRs and then used the Schwartz algorithm [14] to identify their potential long-forms. This simple method can essentially be used without worrying about the type of the text and the language. The automated extraction resulted in many potential long-forms for each short-form. Although others have considered this a word sense disambiguation (WSD) problem and tried to tackle it accordingly [12, 13], we used a naïve approach to select the most frequent long-form among the options. Even after implementing several pre- and post-processing filters to remove false positives, we still ended up with plenty of incorrect short-form long-form pairs. It would be interesting to see if more sophisticated WSD techniques could reduce the false positives and produce better short-form long-form mapping.

In our third approach, we used clustering methods to normalize all word-variations to one. Previous studies on normalizing clinical text mainly focused on spelling corrections [15–18]. Most of these methods are dependent on linguistic resources or domain-specific terminologies, which are hard to come by for non-English languages such as Dutch. Of the two methods we used, the normalization using edit-distance-based clusters resulted in 38% reduction in features whereas only 5% reduction was observed when lemma-based word normalization was applied. A reason for the low reduction using lemma-based clustering could be that the lemmatizer was trained on a non-clinical Dutch lexicon and may not work as good on clinical terms. Another plausible reason might be that the text in IPCI is noisy and since the lemmatization process usually involves a vocabulary and morphological analysis of the words, it may have a strong impact on the performance. Our results show that the feature reduction had a (slightly) positive impact on the classification performance and resulted in improved sensitivity, which is usually important for clinical data sets. It could, therefore, be concluded that for EHR databases with noisy unstructured text, an edit-distance-based clustering approach could prove to be beneficial as it cannot only reduce the feature dimensionality but can also increase sensitivity of the classification tasks.

Mining Electronic Health Records

Case selection is one of the most important steps in observational studies. The algorithms for case selection are commonly known as *case-detection algorithms*. Case-detection algorithms are usually created manually by using structured information such as ICD-9 codes or laboratory values since large amounts of free-text present in most EHRs cannot be easily analyzed [19]. In databases that contain unstructured text, the manually crafted algorithms consist of all possible words and codes that might be relevant. A typical process of case selection involves applying manual algorithms and then manually verifying the potential patients by reviewing their medical records [20]. In Chapter 5, we explored machine-learning methods to generate and evaluate case-detection algorithms to identify children with asthma within the IPCI database. Considering the hierarchical nature of the asthma labels in the study (definite -> probable -> doubtful -> non-asthma), we tackled this as a hierarchical multi-class classification problem [21–23]. We trained two machine-learning classifiers, one to separate definite cases from all other cases and the other

to distinguish probable and doubtful asthma from non-asthma cases. We used a third rule-based classifier to distinguish between probable and doubtful asthma cases.

Although we used both structured information and unstructured free-text to train the classifier, none of the seven rules generated by the algorithm for definite asthma cases contained any diagnosis code. This is particularly interesting since all definite asthma cases required an explicit confirmation by the specialists. The algorithm successfully found such confirmations, which are usually found in the free-text entries of the IPCI database. This highlights the importance of using unstructured free-text in the case selection process and the ability of the automated algorithm to identify such vital information. This is in line with recent findings showing that the use of unstructured text can significantly improve case detection [19]. Typically, researchers are required to write manual case-detection algorithms. These researchers are also required to have good understanding of the database, such as how the information was collected and how it is stored, where to find required information, what coding schemes are used to record diagnosis, symptoms, and drugs, what information can be found in the structured, and what can be expected from the unstructured text. For unstructured text, it is also important to know whether the free-text is free from typographical errors and if non-standard abbreviations and acronyms are used. Poor knowledge of the database often results in sub-optimal manual case-detection algorithms leading to either missing many cases or including many false positives. Errors in the case selection process lead to misclassification and potential biased findings [24, 25]. The machine-learning methods that we used to generate automated case-detection algorithms, showed that they are capable of capturing specific keywords (and combinations) used within a database. For example, the algorithm to classify asthma in the IPCI database contained 'flixotide', 'ventolin', and 'pulmicort' which are drugs for obstructive airway diseases, database-specific code '20' for the department of pediatrics, non-standard abbreviations such as 'inh' for 'inhaler' and 'vag' for 'vesiculair ademgeruis' (vesicular breath sounds). We showed that the performance of the automated case-detection algorithm is on a par with manual annotators [26]. Another advantage of automated case-detection algorithms is that they can allow for more uniform and consistent annotations as compared with several manual annotators.

The handcrafted case-detection algorithms usually have very high sensitivity since they tend to include all possible keywords (code and/or text), but it comes at the expense of a manual review step to wield out the false-positives, which is very time-consuming and labor-intensive. Although there is no general agreement on how much error is acceptable in an automated case selection process [19], an ideal algorithm would have high sensitivity and a high positive predictive value (PPV). In incidence rate studies in particular, where the goal is to find the number of new cases in the population during a given time period, the automated case-detection algorithm should have high sensitivity, i.e., it should not miss many positive cases. However, machine-learning methods are usually optimized for accuracy but not for sensitivity. The quality of the machine-learning methods greatly depends on the training set. Often, there are more negative cases than positive cases in the training set and this imbalance affects the machine-learning process [27]. In Chapter 6, we explored different methods to deal with the training set imbalance in order to achieve high sensitivity. One way of improving algorithm performance is to have a large training set with a high number of positive cases, which is expensive. Increase in sensitivity usually comes

at the cost of a decrease in specificity due to the inclusion of many false positives. We experimented with two methods in order to increase sensitivity of the case-detection algorithms. In the first method, we varied the imbalance ratio between positive and negative cases in the training set using under-sampling and over-sampling techniques. In the second method, we applied cost-sensitive learning techniques with various misclassification costs. We observed about equal performance for both methods, which is in line with previous findings [28]. However, the main challenge with cost-sensitive learning was to find an optimal cost matrix since different classifiers treat cost differently, and it depends on the data set and the classifier used. We used random under- and over-sampling methods to find the best class distribution ratio in the data set in order to achieve high sensitivity without severely jeopardizing specificity. Among these, there was no clear winner as they both performed consistently well. The under-sampling methods are in general preferred since over-sampling methods are prone to overfitting [29]. The question of finding an optimal class distribution ratio has been studied by several researchers in the past [27, 29–31]. A distribution close to the naturally occurring class distribution is reported to achieve good accuracy but a more balanced class distribution tends to maximize Area Under the Curve (AUC). Our experiments showed that the classifiers achieved high sensitivity with not too low specificity when the imbalance ratio (negative cases to positive cases) was below 10. This effect was observed using four different machine-learning algorithms on two very different clinical data sets. Both cost-sensitive learning and sampling methods were able to achieve high sensitivity similar to the manual annotators. Since all potential cases are manually validated during case selection process, such automated case-detection algorithms with high sensitivity can be used as a pre-filter to significantly reduce the workload and save weeks of manual work in an epidemiological study.

The data contained in EHRs are collected for clinical purposes and primarily used for routine medical care. The extensive amount of healthcare information present in EHRs has also allowed researchers to conduct health outcome research, especially to study post-marketing drug effects [32–34]. However, researchers need to be wary of the issues related to using the EHRs for these studies, such as confounding. Confounding occurs when a third variable that is not under investigation, is associated with both the exposure and the outcome of interest. Observational studies need to deal with confounding by design (restriction), matching, or adjustment. Statistical techniques such as a propensity score [35] can be used to address the confounding through matching or adjustment. In Chapter 4, we explored the possibility of using unstructured free-text in the IPCI database to construct propensity score models that would allow to deal with confounding. EHRs comprise much unstructured data that could be used as proxies for potential confounding factors. These factors are difficult to capture from EHRs because the information is not primarily recorded for research purposes. Previous studies on confounding control that use propensity score models focused on including only structured information, such as diagnostic or procedure codes available in claims databases [36–38]. A high-dimensional propensity score (hd-PS) algorithm was proposed to empirically identify large number of relevant covariates with high prevalence [39]. The use of two-word free-text phrases in addition to the structured information has also been positively evaluated in the context of hd-PS models [40].

Our method to construct propensity score models is different since we used all unstructured text without pre-identifying data dimensions as compared to the hd-PS. We considered all unigrams as potential covariates that could enter the propensity score model. We generated two different propensity score models; the first used covariates with the highest frequencies in the cohort and the second used covariates with an association with the outcome. Our results suggest that a high frequency threshold could be used to select covariates since it appears that the generated models are mostly based on a few covariates with high occurrence in the text. However, such a frequency-based covariate selection approach is prone to include covariates that may actually be instrumental variables. If covariates are included that are not true confounders, the variance increases and sometimes a small amount of bias may be introduced [41–43]. To mitigate this, in the second method we included covariates with a significant association with the outcome to the propensity score model. Our results showed that this method provided an improvement in adjustment for confounding. Using only the unigrams was one of the limitations of our study. Some important covariates like 'congestive heart failure' could not be recognized as such; instead they were recognized as individual words 'congestive', 'heart', and 'failure'. This could lead to over- and under-estimation of some covariates. Therefore, we suggest that more efforts should be spend on developing better methods to extract meaningful covariates from the free-text for effective proxy adjustment via propensity scores.

Dutch Clinical Corpus Development

The availability of annotated corpora is essential to train and test automated language-processing systems. The performance of such systems usually depends on the quality and quantity of the annotations.

There was no clinical corpus available in the Dutch language suitable to train and evaluate systems to identify contextual properties. Therefore, we developed a clinical corpus consisting of 7500 textual entries in the Dutch language. Four types of clinical documents are included in the corpus to capture different language use in the Dutch clinical setting. The combination of these texts can be considered a representative selection of the documented medical process in the broadest sense, including the patient's first interactions with the general practitioner, referrals and advanced (imaging) diagnostics in the hospital, and ultimately reporting back to the general practitioner after polyclinic consult or discharge after hospital admission. Developing a high quality corpus depends on two things: a) clear annotation guidelines and b) trained subject matter experts to carry out the annotation work. Our annotation guidelines consisted of clear definitions of each of the contextual properties with examples. Several one-on-one sessions were conducted with the two medically trained annotators to make sure they understood the definitions and the annotation tool [44] that was used for the task. Finding all medical conditions or symptoms first and then identifying three of their contextual properties is a tedious and labor-intensive job. To speed up the corpus development, we limited the annotations to the conditions already identified in the text using our custom Dutch UMLS terms. Therefore, the annotator's task was reduced to only identify their contextual properties. A more extensive term list would have identified many more terms in the clinical text.

We observed that the medically schooled annotators were prone to using information outside the context and make considerations based on prior knowledge concerning the natural course of a disease. For example, on various occasions one annotator labeled a term as 'historical' based on the assumed chronicity of the disease without presence of explicit evidence in the context. Automated methods to recognize properties of clinical terms only from its context can never identify this and may result in biased estimations. Overall, the inter-annotated agreement, measured using Cohen's Kappa [45], was good to very good except for one of the values ('hypothetical') of the temporality property, which was moderate. The annotators often disagreed on the assignment of hypothetical values to terms that were part of a differential diagnosis. An expert, who was familiar with all four types of clinical text, resolved the disagreements between the two annotators. The anonymized Dutch clinical corpus we developed is the first publically available clinical corpus in the Dutch language. It can serve as a useful resource for further algorithm development.

Limitations and future work

In this thesis, we focused on two main steps of the knowledge-discovery process: data preparation and data mining. In data preparation, the efforts were spent primarily on data cleaning and data reduction tasks. Although we successfully adapted an English language algorithm to detect contextual properties of clinical terms in Dutch, there were still some challenges. The ContextD algorithm only considers a sentence as a context. This leads to errors if sentence boundaries are not correctly identified. We used the Apache OpenNLP [46] library to split text into sentences, which is trained on regular natural language text and sometimes failed to correctly identify a proper sentence. We see an opportunity to increase the algorithm performance by using a sentence splitter trained on Dutch EHRs. The ContextD algorithm should also be extended to use additional useful information present outside of the sentence-level context. For example, all clinical concepts identified in the 'patient history' section should be labeled as historical but this information is not available to the algorithm. The ContextD algorithm is based on pre-identified trigger words. The algorithm uses exact string matching to check the presence of a trigger in a sentence. We noticed a few errors due to typos in trigger words. These errors could be reduced by employing a more sophisticated fuzzy matching technique for finding trigger terms in the text.

In one of our studies, we tried to identify potential short-forms (abbreviations and acronyms) and map them to their long-forms within the IPCI database. This proves to be challenging since the database contained many standard and non-standard abbreviations. The Schwartz algorithm [14] often identified many possible long-forms for one short-form. To disambiguate, we simply used frequency information and map the most frequent long-form to the short-form. Normalizing short-forms to long-forms is a challenging task and future work in this domain may focus on investing in building more robust deep learning approaches [47]. We also showed that in IPCI-like databases containing plenty of noisy free-text, an edit-distance based clustering technique could be used to reduce the feature dimensionality. A similar approach using lemmatization did not work well mainly because the Dutch lemmatizer [48] we used was not

trained on medical text. Further use of the Dutch lemmatizer should focus on retraining the in-build models first to achieve better results.

Machine-learning approaches can be used to generate case-detection algorithms automatically by using unstructured and structured information. Feature extraction is essential to apply traditional machine-learning techniques but it requires data and domain knowledge. New machine-learning approaches such as deep learning can be used to learn features from the data set algorithmically in an unsupervised way [49]. Although the use cases in this thesis were focusing on Dutch EHRs, the approaches are more general and can be used with other databases. However, this would require new training sets to build data-specific algorithms. We also showed how the sensitivity of machine-learning methods could be improved by dealing with the imbalance in the training set. We found simple random under-sampling and over-sampling approaches to be beneficial. However, both these approaches were used mutually exclusive to each other. Since both approaches have their own limitations, we suggest that further investigations should be looking into applying a combination of both under-sampling and over-sampling methods at the same time.

Although the data preparation techniques presented in this thesis can greatly benefit the subsequent data-mining step, there are still many opportunities to improve. Quantitative information in the EHRs such as laboratory results, body temperature, and blood pressure measurements could be identified as well. This would first require an identification of the quantitative attribute and second its value from the context. The coded information present in the structured part of the EHRs may also be present in a textual form in the unstructured free-text. Therefore, it is important that such information is used only once to avoid any potential bias. Identifying information overlap between structured and unstructured free-text is a challenging task that requires further efforts. Similarly, more efforts should be made to structure the unstructured free-text in an automated fashion as much as possible. Typically, only the definitive diagnosis information is coded in the EHRs and not the related symptoms. Use of named entity recognition techniques to identify symptoms, leading to a definite diagnosis, can greatly benefit in understanding the natural history of disease [1].

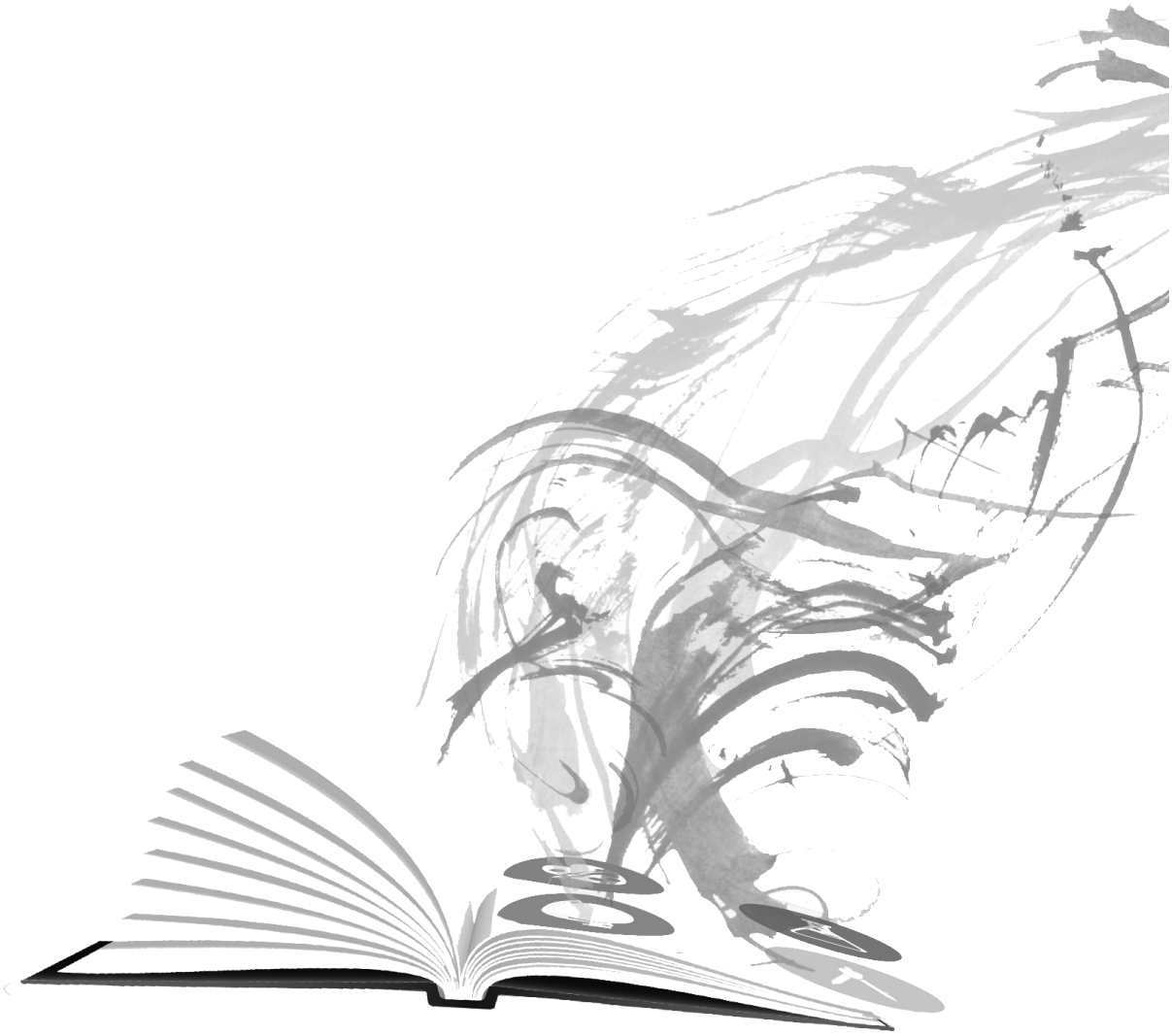
Every EHR database is structurally different from another and that limits the possibilities of combining multiple EHRs database to increase sample size for observational studies. Moreover, due to the differences, methods developed for one EHR database cannot be used straightaway with other databases. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [50] provides a viable framework to convert an EHR database to a common structure to not only combine multiple EHRs but also take advantage of new methods and tools developed by the Observational Health Data Sciences and Informatics (OHDSI) program [51]. However, extracting and converting information from unstructured free-text to the OMOP-CDM is largely unexplored and merits further work.

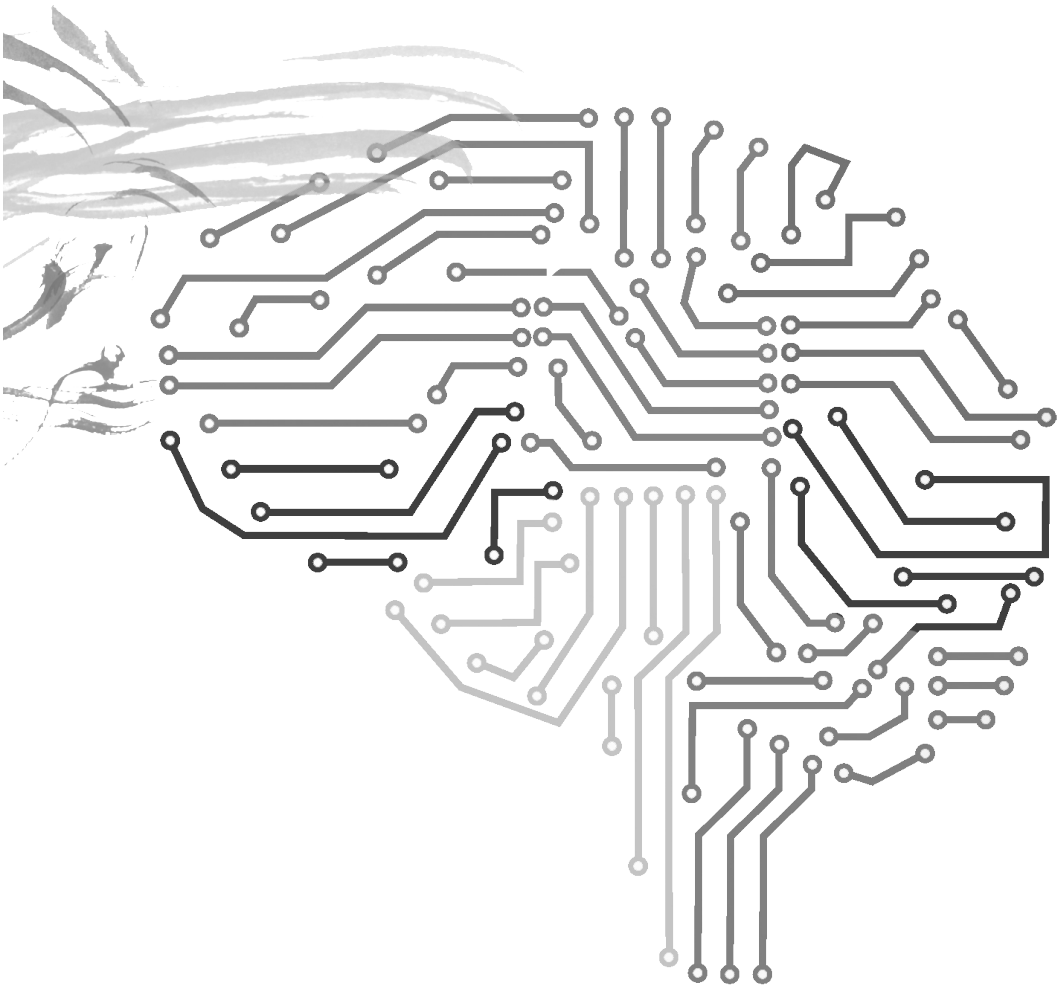
REFERENCES

1. Yadav P, Steinbach M, Kumar V, Simon G: **Mining Electronic Health Records: A Survey**. *ACM Comput Surv* 2016, **1**:1–41.
2. **Integrated Primary Care Information (IPCI) http://www.ipci.nl** [http://www.ipci.nl]
3. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: **A simple algorithm for identifying negated findings and diseases in discharge summaries.** *J Biomed Inform* 2001, **34**:301–310.
4. Skeppstedt M: **Negation detection in Swedish clinical text: An adaption of NegEx to Swedish**. *J Biomed Semantics* 2011, **2**(Suppl 3):S3.
5. Deléger L, Grouin C: **Detecting negation of medical problems in French clinical notes**. In *Proc 2nd ACM SIGHIT Symp Int Heal informatics - IHI '12*. New York, New York, USA: ACM Press; 2012:697–702.
6. Vivaldi VCVSJ, Rodriguez H: **Syntactic methods for negation detection in radiology reports in Spanish**. In *ACL*; 2016:156.
7. Costumero R, Lopez F, Gonzalo-Martín C, Millan M, Menasalvas E: **An approach to detect negation on medical documents in Spanish**. In *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. *Volume 8609 LNAI*; 2014:366–375.
8. Harkema H, Dowling JN, Thornblade T, Chapman WW: **ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports.** *J Biomed Inform* 2009, **42**:839–851.
9. **Google Translate** [http://translate.google.com]
10. Chapman W, Chu D, Dowling J: **ConText: An algorithm for identifying contextual features from clinical text**. In *Proc Work BioNLP 2007 Biol Transl Clin Lang Process*. Prague, Czech Republic: Association for Computational Linguistics; 2007(June):81–88.
11. Schuemie MJ, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup**. In *Proc Second BioCreative Chall Eval Work*; 2007:131–133.
12. Mowery DL, South BR, Christensen L, Leng J, Peltonen L-M, Salanterä S, Suominen H, Martinez D, Velupillai S, Elhadad N, Savova G, Pradhan S, Chapman WW: **Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2**. *J Biomed Semantics* 2016, **7**:43.
13. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H: **A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries.** *AMIA . Annu Symp proceedings AMIA Symp* 2012, **2012**:997–1003.
14. Schwartz AS, Hearst MA: **A simple algorithm for identifying abbreviation definitions in biomedical text.** *Pac Symp Biocomput* 2003:451–462.
15. Patrick J, Sabbagh M, Jain S, Zheng H: **Spelling correction in clinical notes with emphasis on first suggestion accuracy**. In *2nd Work Build Eval Resour Biomed Text Min*; 2010(March):2–8.
16. Crowell J, Zeng Q, Ngo L, Lacroix E-M: **A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries**. *J Am Med Informatics Assoc* 2004, **11**:179–185.
17. Tolentino HD, Matters MD, Walop W, Law B, Tong W, Liu F, Fontelo P, Kohl K, Payne DC: **A UMLS-based spell checker for natural language processing in vaccine safety.** *BMC Med Inform Decis Mak* 2007, **7**:3.
18. Lai KH, Topaz M, Goss FR, Zhou L: **Automated misspelling detection and correction in clinical**

free-text records. *J Biomed Inform* 2015, **55**:188–195.
19. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA: **Extracting information from the text of electronic medical records to improve case detection: a systematic review**. *J Am Med Informatics Assoc* 2016:ocv180.
20. Afzal Z, Engelkes M, Verhamme KMC, Janssens HM, Sturkenboom MCJM, Kors JA, Schuemie MJ: **Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases**. *Pharmacoepidemiol Drug Saf* 2013, **22**.
21. Kiritchenko S, Matwin S, Nock R, Famili AF: **Learning and evaluation in the presence of class hierarchies: Application to text categorization**. *Adv Artif Intell* 2006, **4013**:395–406. [*Lecture Notes in Computer Science*]
22. Metz J, Freitas AA, Monard MC, Cherman EA: **A study on the selection of local training sets for hierarchical classification tasks**. In *Brazilian Natl Meet Artif Intell*. Natal, RN, Brasil: Sociedade Brasileira da Computa - SBC; 2011:572–583.
23. Costa EP, Lorena AC, Carvalho ACPLF, Freitas AA, Holden N: **Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees**. In *Adv Bioinforma Comput Biol*. *Volume 4643*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007:126–137.
24. Nicholson A, Tate AR, Koeling R, Cassell JA: **What does validation of cases in electronic record databases mean? The potential contribution of free text.** *Pharmacoepidemiol Drug Saf* 2011, **20**:321–4.
25. Manuel DG, Rosella LC, Stukel TA: **Importance of accurately identifying disease in studies using electronic health records**. *Br Med J* 2010, **341**:440–443.
26. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors J a: **Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records.** *BMC Med Inform Decis Mak* 2013, **13**:30.
27. Garcia EA: **Learning from Imbalanced Data**. *IEEE Trans Knowl Data Eng* 2009, **21**:1263–1284.
28. McCarthy K, Zabar B, Weiss G: **Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?** In *Proc 1st Int Work Util data Min - UBDM '05*. New York, New York, USA: ACM Press; 2005:69–77.
29. Weiss GM, Provost F: **Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction**. *J Artif Intell Res* 2003, **19**:315–354.
30. Chawla N V: **Data Mining for Imbalanced Datasets: An Overview**. In *Data Min Knowl Discov Handb*. Edited by Maimon O, Rokach L. Boston, MA: Springer US; 2010:875–886.
31. Van Hulse J, Khoshgoftaar TM, Napolitano A: **Experimental perspectives on learning from imbalanced data**. In *Proc 24th Int Conf Mach Learn - ICML '07*. New York, New York, USA: ACM Press; 2007:935–942.
32. Myers L, Stevens J: **Using EHR to Conduct Outcome and Health Services Research**. In *Second Anal Electron Heal Rec*. Cham: Springer International Publishing; 2016:61–70.
33. MIT Critical Data: *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing; 2016.
34. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW: **Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting**. *Pharmacoepidemiol Drug Saf* 2010, **19**:1211–1215.
35. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S: **Confounding control in healthcare database research: challenges and potential approaches**. *Med Care* 2010, **48**:S114–S120.

36. Rubin DB: **Estimating causal effects from large data sets using propensity scores.** *Ann Intern Med* 1997, **127**:757–763.
37. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T: **Variable selection for propensity score models**. *Am J Epidemiol* 2006, **163**:1149–1156.
38. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S: **Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples**. *Am J Epidemiol* 2011, **173**:1404–1413.
39. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA: **High-dimensional propensity score adjustment in studies of treatment effects using health care claims data.** *Epidemiology* 2009, **20**:512–522.
40. Rassen JA, Wahl PM, Angelino E, Seltzer MI, Rosenman MD: **Automated Use of Electronic Health Record Text Data To Improve Validity in Pharmacoepidemiology Studies**. In *Pharmacoepidemiol Drug Saf*. *Volume 22*. NJ USA: WILEY-BLACKWELL; 2013:376.
41. Myers J a., Rassen J a., Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ: **Effects of adjusting for instrumental variables on bias and precision of effect estimates**. *Am J Epidemiol* 2011, **174**:1213–1222.
42. Rassen JA, Schneeweiss S: **Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system**. *Pharmacoepidemiol Drug Saf* 2012, **21**:41–49.
43. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S: **Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses**. *Am J Epidemiol* 2015, **182**:651–659.
44. **brat rapid annotation tool** [http://brat.nlplab.org/]
45. Application I-RR, Kappa I, Yellow-Bellied F, Red-Bellied F, Cooters R: **Cohen â€™ s Kappa**. *Communication* 1960:1–3.
46. **Apache OpenNLP library** [http://opennlp.apache.org/]
47. Wu Y, Xu J, Zhang Y, Xu H: **Clinical abbreviation disambiguation using neural word embeddings**. In *Proc 2015 Work Biomed Nat Lang Process*; 2015:171–176.
48. van den Bosch A, Busser B, Canisius S, Daelemans W: **An efficient memory-based morphosyntactic tagger and parser for Dutch**. In *Sel Pap 17th Comput Linguist Netherlands Meet*. Edited by Eynde F V, Dirix P, Schuurman I, Vandeghinste V. Leuven, Belgium; 2007:99–114.
49. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E: **Deep learning applications and challenges in big data analytics**. *J Big Data* 2015, **2**:1.
50. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE: **Validation of a common data model for active safety surveillance research.** *J Am Med Inform Assoc* 2012, **19**:54–60.
51. **Observational Health Data Sciences and Informatics (OHDSI)** [http://www.ohdsi.org/analytic-tools/]





Summary

This thesis describes the use of several data-mining and data-preparation techniques for automated processing of Dutch electronic health records. All data sets used in this thesis were taken from the Integrated Primary Care Information (IPCI) database, which is a longitudinal collection of EHRs from Dutch general practitioners. We started with implementing and evaluating an algorithm to identify contextual properties of clinical concepts. Next, we looked at ways to normalize abbreviations and textual variations to reduce feature dimensionality. We continued with assessing whether the unstructured information in EHRs can also be used to construct propensity score models to deal with confounding. Later we generated and evaluated case-detection algorithms for case selection, which is an important task in observational studies. Finally, we continued with exploring the options to improve performance of generated case-detection algorithms with focus on the sensitivity. Next, we present a summary of the main findings discussed in this thesis.

In Chapter 2, we presented ContextD, an adaptation of the English ConText algorithm to the Dutch language. The algorithm is able to identify three contextual properties: negation, temporality, and experiencer of the clinical concepts. To adapt the algorithm, we translated all English trigger terms to Dutch and added several general and Dutch EHR-specific enhancements such as negation rules for general practitioners' entries and a regular expression based temporality module. We also developed a Dutch clinical corpus to evaluate ContextD. The performance of the ContextD was better than the original ConText algorithm in identifying negation property but lower for identifying historical properties in discharge letters. The corpus was annotated for three contextual properties and consisted of four different types of Dutch EHRs. The Dutch clinical corpus has been made public and can be used to train other systems for similar tasks.

In Chapter 3, we normalized words in EHRs in order to reduce feature dimensionality. We employed two approaches for normalization. In the first approach, we group textually similar words together using clustering methods, and in the second approach, we identified abbreviations and acronyms and mapped them to their long-forms that are present in the EHRs. We managed to greatly reduce feature dimensionality using a word clustering based normalization approach. We also showed that word normalization resulted in better classification performance, especially in improving sensitivity.

Studies using EHRs often have to deal with confounding, which occurs when a variable that is not under investigation influences the outcome. To deal with confounding, in Chapter 4, we used a large-scale regularized regression to fit two propensity score (PS) models using all structured and unstructured information in the EHR. We generated two different PS models: the first was generated using covariates with the highest frequencies in the cohort and the second was generated using covariates with an association with the outcome. We showed that these PS models provided an improvement in adjustment for confounding. This is useful for database studies that have a large amount of unstructured free-text as in EHRs.

Case detection algorithms are usually created manually and they often use only the structured or coded information in the EHRs, such as ICD-9 codes. In Chapter 5, we used machine-learning methods to generate and evaluate an automated case-detection algorithm that uses both coded information and free-text. The generated algorithm yielded high sensitivity and specificity on identifying asthma cases. Automating case selection by means of auto-generated case-detection algorithms will facilitate large-scale studies from databases.

Machine-learning methods are typically optimized for accuracy but not for sensitivity. In Chapter 6, we explored two methods to handle the imbalance in the training set with focus on improving the sensitivity of the resulting classifiers. On two evaluation sets, we were able to achieve high sensitivity on par with the manual annotators. By tweaking the training set balance of positive and negative examples, we managed to improve sensitivity on the first set by 4% and on the second set by 20%. Highly sensitive case-detection algorithms can be used as a pre-filter to significantly reduce the burden of manual record validation during the case selection process.

SAMENVATTING VAN DE BELANGRIJKSTE BEVINDINGEN

Dit proefschrift beschrijft het gebruik van verschillende data-mining- en bewerkingstechnieken voor de automatische verwerking van Nederlandse patiëntendossiers. Gegevens gebruikt voor dit proefschrift kwamen uit de Integrated Primary Care Information (IPCI) database en het ziekenhuisinformatiesysteem van het Erasmus MC. IPCI is een longitudinale verzameling van elektronische patiëntendossiers (EPDs) van Nederlandse huisartsen. Het onderzoek startte met het bouwen en evalueren van een algoritme om contextuele eigenschappen van klinische concepten te identificeren. Vervolgens onderzochten we manieren om afkortingen en tekstvariaties te normaliseren, met als doel het aantal featuredimensies terug te brengen. We beoordeelden of de ongestructureerde informatie in EPDs ook gebruikt zou kunnen worden om zgn. propensity score modellen te construeren, die onverwachte afwijkingen ('confounding') zouden kunnen behandelen. Later bouwden en evalueerden we algoritmes om casussen te herkennen en selecteren, een belangrijke taak in observationele studies. Tenslotte exploreerden we de opties om de prestaties van casus-selectie-algoritmes te verbeteren, met de focus op gevoeligheid. Hieronder volgt een korte samenvatting van de belangrijkste bevindingen in dit proefschrift.

In hoofdstuk 2 presenteerden we ContextD, een aanpassing van het Engelse ConText algoritme aan de Nederlandse taal. Het algoritme is in staat om drie contekstuele eigenschappen te identificeren: negatie, temporaliteit en onderwerp van de klinische concepten. Om het algoritme aan te passen vertaalden we alle Engelse 'trigger'-termen naar het Nederlands, en voegden verschillende algemene en EPD-specifieke verbeteringen toe, zoals negatie-regels voor notities van huisartsen, en een module om tijdsbepalingen te herkennen, gebaseerd op reguliere expressies. We ontwikkelden ook een Nederlands klinisch corpus om ContextD te evalueren. De prestaties van ContextD waren beter dan die van het originele ConText algoritme met betrekking tot de herkenning van negatie, maar minder goed in het identificeren van temporaliteit-gegevens in ontslagbrieven. Het corpus is geannoteerd met drie contekstuele eigenschappen, en bestaat uit vier verschillende types Nederlandse EPDs. Het Nederlands klinisch corpus is gepubliceerd, en kan worden benut om andere systemen voor gelijksoortige doelen te trainen.

In hoofdstuk 3 normaliseerden we woorden in EPDs met de bedoeling het aantal feature-dimensies terug te brengen. We kozen twee benaderingen van normalisatie. In de eerste benadering groepeerden we tekstueel gelijkende woorden met clustering-methodieken, en in de tweede benadering identificeerden we afkortingen en acroniemen om ze te koppelen aan voluit-geschreven woorden in EPDs. Op die manier konden we het aantal dimensies aanzienlijk terugbrengen. Tegelijkertijd toonden we aan dat woordnormalisatie resulteert in betere prestaties van de classificatie, vooral in het verbeteren van sensitiviteit.

Observationele studies hebben vaak te maken met confounding, omdat artsen handelen met de prognose van de patiënt in gedachten. Deze prognose kan leiden tot het fenomeen dat een geneesmiddel wordt voorgeschreven vanwege een prognose en dat die prognose is gerelateerd aan de uitkomst. Om dit probleem te adresseren, hebben we in hoofdstuk 4 twee propensity score (PS) modellen gemaakt, waarbij we zowel gestructureerde als ongestructureerde

informatie in het EPD benutten. We genereerden twee verschillende PS modellen: de eerste op basis van covariabelen met de hoogste frequenties in het cohort, en de tweede op basis van covariabelen geassocieerd met de uitkomsten. We toonden aan dat deze PS modellen een verbetering geven in het corrigeren van confounding. Dit is nuttig voor database-studies met een grote hoeveelheid ongestructureerde informatie, zoals in EPDs.

Algoritmes om ziekte te identificeren in EPDs worden vaak handmatig gemaakt, wat de reproduceerbaarheid en schaalbaarheid negatief beïnvloedt. In hoofdstuk 5 gebruikten we machine-learning methodes om een geautomatiseerd casus-herkenningsalgoritme te ontwikkelen en evalueren. Dit algoritme gebruikte zowel gecodeerde informatie als vrije tekst. Het algoritme presteerde met hoge sensitiviteit en specificiteit op het identificeren van astma-casussen. Automatisering van casus-selectie door machine-gegenereerde algoritmes zal grootschalige database-studies faciliteren.

Machine-learning methods worden gewoonlijk geoptimaliseerd voor accuratesse, maar niet voor sensitiviteit. In hoofdstuk 6 exploreerden we twee manieren om de disbalans in de training set aan te pakken, met als doel de sensitiviteit van de resulterende classificatiealgoritmes te verbeteren. Op twee evaluatie-sets bleek het mogelijk om een sensitiviteit te bereiken die vergelijkbaar is met handmatige annotatie. Door de balans van positieve en negatieve voorbeelden aan te passen, slaagden we erin om sensitiviteit op de eerste set met 4%, en op de tweede set met 20% te verhogen. Casus-herkenningsalgoritmes met hoge sensitiviteit kunnen gebruikt worden bij wijze van voorselectie om de werklast van handmatige dossiervalidaties aanzienlijk te verminderen.

ACKNOWLEDGEMENTS

Alhamdulillah, writing this note of thanks marks the end of this journey successfully. As with most PhD journeys, this journey is also filled with ups and downs, happiness, satisfactions and occasional frustrations. It has been a period of valuable learning for me, both on professional and personal level. This journey could not have been possible without the love and understanding of my family and the support of several colleagues and friends around.

First of all, I would like to thank my promotor Prof. dr. Miriam Sturkenboom for giving me this opportunity. Miriam, I still remember the day I came to Rotterdam for an interview and somehow we ended up talking about food and cooking. I don't know how much this contributed to my selection but I am glad you took a chance on me. I thank you sincerely for your encouragement, guidance and for always showing confidence in my work.

To my co-promotors Dr. Jan A. Kors and Dr. Martijn J. Schuemie, thank you! Jan, you have been an amazing supervisor all these years. Your attention to detail is second to none. I have learned a lot from you, especially on how to be thorough, critical, and about never compromising on quality. I cannot thank you enough for your untiring reviews of my manuscripts to make them what they are and always being there to guide and help. Martijn, your ability to quickly understand a problem and suggest an answer is remarkable. I was lucky to have you as my supervisor. Thank you for always finding time for me even on your busiest days. I am thankful for all the help and guidance you have provided me with.

I would also like to thank inner doctoral committee, Prof.dr. J.A. Hazelzet, Prof.dr. A. van den Bosch, and Prof.dr. A. Abu-Hannah for their valuable time and encouraging feedback.

Prof. Johan van der Lei, thank you for creating such a friendly environment in the department and ensuring that all the staff and PhD students are well taken care of. All the colleagues at the Department of Medical Informatics, thank you for making these years memorable. Special thanks to all the secretariat staff: Desiree, Tineke, Carmen, Petra, and of course Sander. Thank you for always being there to help with smiling faces. My wonderful colleagues in the Biosemantics group: Saber, Erik, Rein, Ewoud, Kang, Benus, Wytze, Chinh, Kristina, Solène, and Herman. How pleasurable it was to work in the same group as you. Saber, what an incredible friend you have become on and off work over the years. It was great fun to share the room with you in the old Erasmus building and then again in the new building. I like how you always end up agreeing with me whenever we have an argument ☺. Thank you for always being there! Many thanks to Bahar as well for inviting me over many times and offering some delicious food. Ewoud, I have not seen a medical doctor doing NLP or machine learning better than you. I will remember all the laughs we had together and thank you for your valuable contribution towards ContextD publication. Erik, I always admired your quick thinking and practical approach towards research. Thank you for being so approachable and always offering to help. Benus, thank you for all the nice chit chats during our coffee breaks. David and Tiago, it was fun to have you guys over in our group from Portugal for couple of months. Osemeke and Rene, thank you for the joyful company in Montreal. The morning banter between you two was one of the highlights of the trip. Gwen, working with you was always fun. You were a good company

in Montreal and Taipei, and of course how can I forget the Hello Kitty ride. Kartini, we did not get to spend a lot of time together, but I have always enjoyed your company and the discussions we had. Peter Rijnbeek, thank you for kick-starting my research career with that great first session on imbalance data and all the reading material. Peter Moorman, I will remember those cricket related discussions we used to have next to the coffee machine. Thank you to all other colleagues in the IPCI and BIGR group.

Andreas Engelhard and Solane Degenaar, many thanks for taking care of all the administrative and visa related issues during my stay at the Erasmus MC. Mees, thank you for your help with the technical stuff over the years. Marcel, thank you for answering all my IPCI related questions.

I would also like to thank all my coauthors, and those with whom I have collaborated with over the years. Thank you all for your contributions.

Dr. Michelle Gregory and Dr. Marius Doornenbal, thank you for your support and understanding during the final writing part of my dissertation. Marius, thank you for those countless reminders that I have to finish my PhD. I guess you have one less thing to worry about now ☺.

I would like to thank all my friends for their wonderful company and support over the years. My stay in the Netherlands would not have been this great without all of you. Special thanks goes to Imran and Suleman for providing good company first in Germany and then later in the Netherlands.

This, and any other achievements in my life, would not have been possible without the endless love, support, and prayers of my family. My parents have always been a source of great inspiration for me. I thank my mother for always encouraging me to go forward and higher in my life. I hope I have made you proud. To my father, losing you was the hardest thing I have ever had to go through in my life. I wish you were here today to witness this. I thank my brothers Umair and Uzair for taking care of our mother in these difficult times. Special thanks to my younger brother Uzair, it might not have been possible for me to stay here and finish my PhD if it was not for his unconditional support and taking special care for our mother.

Last but not the least, I would like to express my appreciation to my wonderful wife. Farah, thank you for your love and support. Thank you for bearing with me and being always there. Thank you for tolerating my laziness and absences. I cannot thank you enough for everything. Maaez and Irha, you two are the biggest bundles of joys in my life.

There are simply too many wonderful people I would like to acknowledge. Please forgive me if I failed to mention your name. Thank you everyone. I am happy that our paths crossed!

I would like to finish my acknowledgement by saying Alhamdulillah!

Zubair

Rotterdam, 2018

PHD PORTFOLIO

Name:	Muhammad Zubair Afzal
Promotor:	Prof. M.C.J.M. Sturkenboom
Copromotor:	Dr.ir. J.A. Kors
	Dr. M.J. Schuemie
Affiliation:	Erasmus University Medical Center
Department:	Medical Informatics

PHD TRAINING

2010	Process and Data Mining, SIKS/LOIS, University of Eindhoven
2010	Dutch-Belgian Information Retrieval Workshop, Radboud University, Nijmegen
2010	Biostatistics for Clinicians, Erasmus University, Rotterdam
2010	Biomedical and Scientific English Writing and Communication, Erasmus University Medical Center, Rotterdam
2011	Databases and Data Mining, Leiden Institute of Advanced Computer Science, University of Leiden
2011	Scientific Presentations, Erasmus University Medical Center, Rotterdam

ORAL PRESENTATIONS

Concept Recognition in French Biomedical Text Using Automatic Translation

- Best-of-labs: eHealth - Conference and Labs of the Evaluation Forum (CLEF), Évora, Portugal 2016
- Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France 2015

Linguistic variability and clinical terminology in a large Dutch general practitioners database

- *International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Taipei, Taiwan 2014*

Generating and evaluating a propensity score model using textual features from electronic medical records

- *International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Taipei, Taiwan 2014*

Automatic generation of a case-detection algorithm for hepatobiliary disease using machine learning on free-text electronic health records

- *International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Barcelona, Spain 2012*

POSTER PRESENTATIONS

Identifying drug-safety signals in electronic health records: An evaluation of automated case detection algorithms with different sensitivity and specificity

- *International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Montreal, Canada 2013*

MEMBERSHIP

International Society of Pharmacoepidemiology, 2012-2014

AWARDS

- **Stanley A. Edlavitch Award**
  International Scociety of Pharmacoepidemiology, Barcelona, Spain, 2012

- **First place** – eHealth Evaluation Lab – Clinical Named Entity Recognition
  Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France 2015

- **First place** – eHealth Evaluation Lab – Clinical Named Entity Recognition
  Conference and Labs of the Evaluation Forum (CLEF), Évora, Portugal 2016

- **First place** – eHealth Evaluation Lab – ICD10 Coding of French Death Certificates
  Conference and Labs of the Evaluation Forum (CLEF), Évora, Portugal 2016

- **Second place** – CHEMDNER-Patents – Chemical passage detection and text classification
  BioCreative V Challenge and Workshop, Sevilla, Spain 2015

- **Third place** – Concepts, Assertions, and Relations in Clinical Text
  I2B2/VA Challenges in Natural Language Processing for Clinical Data, Washington, DC, USA 2010

LIST OF PUBLICATIONS

1. **Zubair Afzal**, Gwen MC Masclee, Miriam CJM Sturkenboom, Jan A. Kors, and Martijn J. Schuemie. *Generating and evaluating a propensity model using textual features from electronic medical records. (submitted)*
2. **Zubair Afzal**, Miriam CJM Sturkenboom, Jan A. Kors, and Martijn J. Schuemie. *Reducing feature dimensionality by normalizing text in electronic health records. (submitted)*
3. Subhradeep Kayal, **Zubair Afzal**, George Tsatsaronis, Sophia Katrenko, Pascal Coupet, Marius Doornenbal, and Michelle Gregory. *Tagging funding agencies and grants in scientific articles using sequential learning models.* BioNLP 2017, 216-221
4. **Zubair Afzal,** George Tsatsaronis, Marius Doornenbal, Pascal Coupet, and Michelle Gregory. *Learning domain labels using conceptual fingerprints: An in-use case study in the Neurology domain.* In European Knowledge Acquisition Workshop, Springer Cham 2016, 731-745
5. **Zubair Afzal**, Saber A. Akhondi, Herman HHBM van Haagen, Erik M. Van Mulligen, and Jan A. Kors. *Concept recognition in French biomedical text using automatic translation.* International Conference of the Cross-Language Evaluation Forum for European Languages, 2016, 162-173
6. Saber A. Akhondi, Ewoud Pons, **Zubair Afzal**, Herman van Haagen, Benedikt FH Becker, Kristina M Hettne, Erik M van Mulligen, Jan A Kors. *Chemical entity recognition in patents by combining dictionary-based and statistical approaches.* Database 2016
7. Erik M. van Mulligen, **Zubair Afzal**, Saber A. Akhondi, Dang Vo, and Jan A. Kors. *Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts.* In Conference and Labs of the Evaluation Forum (CLEF) 2016, 171-178
8. Ewoud Pons, Benedikt FH Becker, Saber A. Akhondi, **Zubair Afzal**, Erik M Van Mulligen, and Jan A Kors. *Extraction of chemical-induced diseases using prior knowledge and textual information.* Database 2016
9. **Zubair Afzal**, Saber A Akhondi, Herman van Haagen, Erik M van Mulligen, and Jan A Kors. *Biomedical concept recognition in French text using automatic translation of English terms.* In Conference and Labs of the Evaluation Forum (CLEF) 2015
10. Ewoud Pons, Benedikt Becker, Saber A. Akhondi, **Zubair Afzal**, Erick M. van Mulligen, and Jan A. Kors. *RELigator: Chemical-disease relation extraction using prior knowledge and textual information.* In Proceedings of the Fifth BioCreative Challenge Evaluation Workshop 2015, 247-253
11. Saber A. Akhondi, Ewoud Pons, **Zubair Afzal**, Herman van Haagen, Benedikt Becker, Kristina M. Hettne, Erik M. van Mulligen, and Jan A. Kors. *Patent mining: combining dictionary-based and machine-learning approaches.* In Proceedings of the Fifth BioCreative Challenge Evaluation Workshop 2015, 102-109
12. **Zubair Afzal**, Ewoud Pons, Ning Kang, Miriam CJM Sturkenboom, Martijn J Schuemie, and Jan A Kors. *ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus.* BMC bioinformatics 2014, 15(1):373
13. **Zubair Afzal**, Martijn J Schuemie, Miriam CJM Sturkenboom, and Jan A Kors. *Linguistic variability and clinical terminology in a large Dutch general practitioners database*. Pharmacoepidemiology and Drug Safety 2014, 23:327

14. **Zubair Afzal**, Gwen Mc Masclee, Miriam CJM Sturkenboom, Jan A Kors, and Martijn J Schuemie. *Generating and evaluating a propensity score Model using textual features from electronic medical records.* Pharmacoepidemiology and Drug Safety 2014, 23:328
15. Ning Kang, Bharat Singh, Chinh Bui, **Zubair Afzal**, Erik M van Mulligen, and Jan A Kors. *Knowledge-based extraction of adverse drug events from biomedical text.* BMC bioinformatics 2014, 15(1):64
16. **Zubair Afzal**, Jan A Kors, Miriam CJM Sturkenboom, and Martijn J Schuemie. *Identifying drug-safety signals in electronic health records: An evaluation of automated case-detection algorithms with different sensitivity and specificity. Pharmacoepidemiology and Drug Safety 2013,22(s1):285*
17. **Zubair Afzal**, Martijn J Schuemie, Jan C. van Blijderveen, Elif F. Sen, Miriam CJM Sturkenboom, and Jan A. Kors. *Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records.* BMC medical informatics and decision making 2013, 13(1)
18. **Zubair Afzal**, Marjolein Engelkes, Katia Verhamme, Hettie Janssens, Miriam CJM Sturkenboom, Jan A. Kors, and Martijn J. Schuemie. *Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases*. Pharmacoepidemioly and Drug Safety 2013, 22
19. Marjolein Engelkes, Hettie Janssens**, Zubair Afzal**, Johan de Jongste, Miriam Sturkenboom, and Katia Verhamme. *Prescription patterns and treatment adherence of asthma controller therapy in children in a Dutch primary care database.* European Respiratory Journal. 2013:42(s57)
20. **Zubair Afzal**, Martijn J. Schuemie, Emine Sen, Geert W't Jong, Mariam CJM Sturkenboom, and Jan A. Kors. *Automatic generation of a case-detection algorithm for hepatobiliary disease using machine learning on free-text electronic health records.* Pharmacoepidemiology and Drug Safety, 2012(21)
21. Ning Kang, **Zubair Afzal**, Bharat Singh, Erik M. van Mulligen, and Jan A. Kors. *Using an ensemble system to improve concept extraction from clinical records.* Journal of Biomedical Informatics, 2012:45(3):423-428
22. Ning Kang, Rogier Barendse, **Zubair Afzal**, Bharat Singh, Martijn J. Schuemie, Erik M. van Mulligen, and Jan A. Kors. *Erasmus MC Approaches to the i2b2 challenge. Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data*, 2010
23. Ning Kang, Rogier Barendse, **Zubair Afzal**, Bharat Singh, Martijn J. Schuemie, Erik M. van Mulligen, and Jan A. Kors. *A concept annotation system for clinical records. Semantic Web Applications and Tools for Life Sciences* (SWAT4LS 2010), 2010

ABOUT THE AUTHOR

Muhammad Zubair Afzal was born on July 26th, 1979 in Rawalpindi, Pakistan. He received his Bachelor degree in Computer Science (Honors) from the Hamdard University, Pakistan. He attained his Master degree in Software Systems Engineering with a specialization in Information Systems, Data Mining, and Data Exploration in 2006 from Aachen University of Technology, RWTH-Aachen, Germany. His research thesis was carried out at Fraunhofer Institute of Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany in collaboration with Deutsche Welle and Westdeutschen Rundfunks (WDR). In September 2006, he joined the subgroup of Discrete Algebra and Geometry of the Department of Mathematics and Computer Science at the Eindhoven University of Technology, The Netherlands as a software engineer where he was involved in developing and incorporating interactive mathematics applications to learning management systems.

In 2007, he joined Software Technology program of 3TU School for Technological Design and attained a Professional Doctorate in Engineeging (PDEng) degree from the Eindhoven University of Technology, The Netherlands. During his PDEng, he carried out several industrial projects of companies like ViNotion, NXP Semiconductors, FEI Company, CERN, and TechUnited. From December 2008 till September 2009 he worked at Philips Healthcare on designing and developing a prototype application to enhance collaborations between clinicians during Tumor Board meetings.

In October 2009, he started to work as a scientific researcher and a PhD candidate at the Biosemantics group, the department of Medical Informatics, Erasmus University Medical Center, Rotterdam. Zubair was supervised by Prof. Dr. Miriam Sturkenboom, Dr. Jan A. Kors, and Dr. Martijn J. Schuemie. His research was aimed at using text mining and natural language processing to support knowledge discovery from electronic health records. His research work have been submitted or published in peer-reviewed scientific journals. During his PhD time, he also actively participated in several community challenges.

As of March 2016, he is working for Elsevier as a Senior NLP Scientist where he is applying state-of-the-art NLP and machine learning techniques to extract information useful for large commercial and research communities.