Learning Predictive Models from Electronic Health Records
Jing Zhao

Academic dissertation for the Degree of Doctor of Philosophy in Computer and Systems Sciences at Stockholm University to be publicly defended on Thursday 2 March 2017 at 13.00 in Lilla hörsalen, NOD-huset, Borgarfjordsgatan 12.

Abstract
The ongoing digitization of healthcare, which has been much accelerated by the widespread adoption of electronic health records, generates unprecedented amounts of clinical data in a readily computable form. This, in turn, affords great opportunities for making meaningful secondary use of clinical data in the endeavor to improve healthcare, as well as to support epidemiology and medical research. To that end, there is a need for techniques capable of effectively and efficiently analyzing large amounts of clinical data. While machine learning provides the necessary tools, learning effective predictive models from electronic health records comes with many challenges due to the complexity of the data. Electronic health records contain heterogeneous and longitudinal data that jointly provides a rich perspective of patient trajectories in the healthcare process. The diverse characteristics of the data need to be properly accounted for when learning predictive models from clinical data. However, how best to represent healthcare data for predictive modeling has been insufficiently studied. This thesis addresses several of the technical challenges involved in learning effective predictive models from electronic health records.

Methods are developed to address the challenges of (i) representing heterogeneous types of data, (ii) leveraging the concept hierarchy of clinical codes, and (iii) modeling the temporality of clinical events. The proposed methods are evaluated empirically in the context of detecting adverse drug events in electronic health records. Various representations of each type of data that account for its unique characteristics are investigated and it is shown that combining multiple representations yields improved predictive performance. It is also demonstrated how the information embedded in the concept hierarchy of clinical codes can be exploited, both for creating enriched feature spaces and for decomposing the predictive task. Moreover, incorporating temporal information leads to more effective predictive models by distinguishing between event occurrences in the patient history. Both single-point representations, using pre-assigned or learned temporal weights, and multivariate time series representations are shown to be more informative than representations in which temporality is ignored. Effective methods for representing heterogeneous and longitudinal data are key for enhancing and truly enabling meaningful secondary use of electronic health records through large-scale analysis of clinical data.

Keywords: Data Science, Machine Learning, Predictive Modeling, Data Representation, Health Informatics, Electronic Health Records.

Stockholm 2017
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-137936

ISBN 978-91-7649-682-4
ISSN 1101-8526

Department of Computer and Systems Sciences
Stockholm University, 164 07 Kista
Learning Predictive Models from Electronic Health Records

Jing Zhao
ABSTRACT

The ongoing digitization of healthcare, which has been much accelerated by the widespread adoption of electronic health records, generates unprecedented amounts of clinical data in a readily computable form. This, in turn, affords great opportunities for making meaningful secondary use of clinical data in the endeavor to improve healthcare, as well as to support epidemiology and medical research. To that end, there is a need for techniques capable of effectively and efficiently analyzing large amounts of clinical data. While machine learning provides the necessary tools, learning effective predictive models from electronic health records comes with many challenges due to the complexity of the data. Electronic health records contain heterogeneous and longitudinal data that jointly provides a rich perspective of patient trajectories in the healthcare process. The diverse characteristics of the data need to be properly accounted for when learning predictive models from clinical data. However, how best to represent healthcare data for predictive modeling has been insufficiently studied. This thesis addresses several of the technical challenges involved in learning effective predictive models from electronic health records.

Methods are developed to address the challenges of (i) representing heterogeneous types of data, (ii) leveraging the concept hierarchy of clinical codes, and (iii) modeling the temporality of clinical events. The proposed methods are evaluated empirically in the context of detecting adverse drug events in electronic health records. Various representations of each type of data that account for its unique characteristics are investigated and it is shown that combining multiple representations yields improved predictive performance. It is also demonstrated how the information embedded in the concept hierarchy of clinical codes can be exploited, both for creating enriched feature spaces and for decomposing the predictive task. Moreover, incorporating temporal information leads to more effective predictive models by distinguishing between event occurrences in the patient history. Both single-point representations, using pre-assigned or learned temporal weights, and multivariate time series representations are shown to be more informative than representations in which temporality is ignored. Effective methods for representing heterogeneous and longitudinal data are key for enhancing and truly enabling meaningful secondary use of electronic health records through large-scale analysis of clinical data.
SAMMANFATTNING

Den pågående digitaliseringen av hälso- och sjukvården, kraftigt påskyndad av det utbredda införandet av elektroniska patientjournaler, genererar enorma mängder data i en lättillgänglig, beräkningsbar form. Detta ger i sin tur goda möjligheter till betydelsefull sekundär användning av kliniska data i strävan att förbättra hälso- och sjukvården, men även för att stödja epidemiologi och medicinsk forskning. För det syftet behövs tekniker för att analysera stora mängder kliniska data på ett både snabbt och effektivt sätt. Medan maskininlärning ger tillgång till de nödvändiga verktygen finns det, på grund av datans komplexitet, många utmaningar med inlärning av effektiva prediktionsmodeller från elektroniska patientjournaler.


This thesis is based on the following original publications, whose reprints were made with kind permission from the publishers. They are referred to in the text by their Roman numerals.


RELATED PAPERS

The following papers also contribute to this thesis, but are not included.


ACKNOWLEDGEMENTS

Five years ago, as I was finishing up my master studies in Groningen and wondering what to do next, I happened to receive an email about the newly funded DADEL project at DSV. This instantly enlightened me! Having studied both information engineering and epidemiology, the project would provide the perfect platform for me to build upon my interdisciplinary background. I immediately started drafting my application to become a doctoral student in DADEL.

Now, at the end of my doctoral studies, I would like to express my deepest gratitude to the project leader, and also one of my supervisors, Henrik Boström. Thank you for giving me the opportunity to work in DADEL and for believing in me from the beginning. You introduced me to the world of machine learning and strengthened my critical thinking as a young researcher. I appreciate the many discussions we had on various research issues. I would like to express the same gratitude to my other supervisor, Lars Asker, who has always been there to support me. Thank you for all your kind help during these years, whenever needed and whatever the issue. Your shared interest in healthcare has been a great source of inspiration.

I would like to express my sincerest appreciation to the data mining group at DSV for both professional support and pleasant lunch company over the years: Henrik, Lars, Isak, Panos, Tony and Ram. I would also like to thank all the co-authors with whom I have collaborated with pleasure – without you, this thesis would not exist in its current shape. Thanks also to all the other PhD students and colleagues at DSV – it has been a great relief never having to feel alone.

Finally, I cannot over-emphasize the importance of my dear family and friends, especially my parents, 张慧灵 and 赵耀军, for their generous love and support. Aron: classmate, colleague, co-author, travel companion, best friend and husband – thank you for every role that you play in my life! You have been the greatest and most amazing surprise of my doctoral studies. Your contribution to this thesis cannot go unnoticed, not only for collaborating on many of the papers, but also for constantly encouraging, inspiring and loving me throughout this journey. Finally, a special thank you to our coming baby for writing this thesis together with me!

Jing Zhao
Stockholm, January 2017
# Table of Contents

**Abstract**

**Sammanfattning**

**List of Publications**

**Acknowledgements**

1. **Introduction**
   1.1 Background
   1.2 Problem
   1.3 Research Questions
   1.4 Contributions
   1.5 Disposition

2. **Electronic Health Records**
   2.1 Data Characteristics
   2.2 Meaningful Secondary Use
   2.3 Challenges and Pitfalls

3. **Predictive Modeling**
   3.1 Machine Learning
   3.2 Data Representation
   3.3 Related Research

4. **Methodology**
   4.1 Research Strategy
   4.2 Evaluation
     4.2.1 Experimental setups
     4.2.2 Learning algorithms
     4.2.3 Performance metrics
     4.2.4 Hypothesis testing
   4.3 Ethical Issues
5 EMPIRICAL INVESTIGATIONS
   5.1 Adverse Drug Event Detection .................................. 42
   5.2 Stockholm EPR Corpus ............................................ 43
   5.3 Study Designs ........................................................ 45
   5.4 Main Findings ........................................................ 46
       5.4.1 Representing heterogeneous types of data ............. 47
       5.4.2 Leveraging the concept hierarchy of clinical codes .... 49
       5.4.3 Incorporating temporal information ...................... 52

6 CONCLUDING REMARKS .................................................... 57
   6.1 Discussion ........................................................... 58
   6.2 Recapitulation and Conclusions .................................. 66
   6.3 Future Directions ................................................... 67

REFERENCES ................................................................. 69
CHAPTER 1

INTRODUCTION

This chapter provides a brief background to how predictive modeling can be used to facilitate the secondary use of electronic health records, which motivates the research presented in the thesis. The central question of data representation is highlighted, along with the specific challenges that concern the analysis of healthcare data. The formulated research questions stem from these challenges and, in turn, provide motivation for the conducted studies. The main contributions are then summarized, followed by a description of the disposition of the remainder of the thesis.
1.1 BACKGROUND

Advances in computer-based medical systems have resulted in the continuous generation of unprecedented amounts of clinical data. The widespread adoption of electronic health record (EHR) systems, in particular, has greatly facilitated the secondary use of clinical data for research purposes [Häyrinen et al., 2008]. Today, in the era of big data, it is widely acknowledged that analyzing this data may have a significant impact on improving healthcare by supporting clinical decision making, preventing avoidable adverse clinical outcomes, identifying high-risk patients, automatically assigning diagnosis codes, detecting diseases at an early stage, and improving statistics on clinical outcomes – just to name a few promising examples. The ability to analyze clinical data on a large scale has the potential to reward us tremendously, not only by effectively allocating medical resources and drastically reducing healthcare costs, but also by enhancing treatment efficacy and patient safety [Pagliari et al., 2007].

Electronic health records are valuable as they contain large-scale observations of patient conditions, treatments and their effects over time. The digitization of healthcare documentation allows for the extraction of phenotypic profiles of patients that can be exploited for providing clinical decision support, as well as for facilitating medical research and epidemiological activities [Jensen et al., 2012]. Electronic health records are heterogeneous and longitudinal, comprising clinical information of various types of data recorded over time. With respect to heterogeneity, a distinction is sometimes made between structured and unstructured EHR data. Structured EHR data typically refers to information such as basic demographic data, drug prescriptions, diagnosis codes and laboratory tests, while unstructured EHR data includes narrative free-text notes and medical images. There is, however, also a marked heterogeneity within these broad categories of information. These heterogeneous types of data record different information and capture various perspectives of patient trajectories in the healthcare process, often complementing each other such that a richer view can be obtained through integration. In order to exploit successfully the massive amounts of longitudinal and heterogeneous data that are stored in EHRs, there are high demands, with respect to both efficiency and effectiveness, on the analytical methods used. As a result, machine learning has emerged to offer promising solutions for large-scale analysis of EHR data [Savage, 2012].
Using machine learning to analyze EHR data typically involves building predictive models from historical data that can then be used to predict some outcome of interest on new data. Predictive models have been used to estimate the risk of patients developing a certain disease, such as heart failure [Oztekin et al., 2009], chronic obstructive pulmonary disease [Himes et al., 2009], respiratory infection [DeLisle et al., 2010], urinary and bowel incontinence [Westra et al., 2011], hypertension [Sun et al., 2014], chronic kidney disease [Perotte et al., 2015], and rheumatoid arthritis [Zhou et al., 2016]. Predictive modeling relies on machine learning algorithms to transform input data (e.g., health records) to some desired output variable (e.g., the risk of developing a certain disease). One key component in this transformation is the predictors describing the characteristics of the input data and hence carrying information from which patterns may be identified. Predictors, also known as features in machine learning, can be any information that is related to the input data. As a result, high-quality features, in terms of relevance and informativeness, are key to building effective predictive models. For example, a patient can be described by different types of information; however, the blood pressure level is likely to be a more relevant feature for estimating the risk of getting heart failure than, say, hair color. Therefore, one of the first steps of learning predictive models from EHR data is to identify relevant features for the targeted task. In machine learning research, this is referred to as data representation or feature engineering [Alpaydin, 2014].

The issue of data representation can be approached in various ways, primarily depending on the potential involvement of domain experts. One traditional, heuristic approach is to create rules for extracting features. For example, a clinician may handcraft rules such as “diastolic blood pressure > 80 mmHg AND systolic blood pressure > 120 mmHg → hypertension”. This approach, however, relies heavily on domain experts and can therefore be considerably time consuming and costly when modeling large amounts of data. An alternative approach can be said to be data-driven and does not require the involvement of human domain experts. In this approach, one may, for instance, create features globally for the entire dataset without using domain knowledge to create different representations for different types of information; alternatively, one may also let the learning algorithm itself identify useful features from more or less “raw” data [Nikravesh et al., 2006]. In reality, these two approaches to data representation are often combined in some way — exploiting both domain knowledge and data-driven feature engineering techniques — in order to achieve the best possible representation. In this thesis, data-driven approaches to data representation
are primarily explored due to the aforementioned benefits, in some cases with the support of domain knowledge. The considered approaches to representing heterogeneous and longitudinal EHR data include globally defined features, but also involve components of learning from data.

1.2 PROBLEM

Data representation is typically emphasized when modeling unstructured EHR data such as narrative notes [Meystre et al., 2008], but is often neglected when dealing with structured EHR data. Most studies, in which predictive models are learned from structured EHR data, tend to adopt straightforward representations and focus instead on the choice of learning algorithm. A straightforward representation often ignores the complexity of the data or merely relies on domain knowledge from clinical experts. By relying on such representations, a large amount of the valuable information that EHRs contain is unfortunately wasted. For example, in [Chazard, 2011], drugs and diagnoses are extracted from EHRs and represented as binary features with only their existence taken into account; additional information, such as the temporal ordering of drugs and diagnoses, remains untapped. Some studies use a rule-based algorithm to extract features, where the rules are typically predefined by clinical experts. However, such algorithms typically only allow a limited number of features to be extracted, leaving the vast majority of the data untouched. For example, in [Perotte et al., 2015], where EHR data is analyzed for predicting the risk of chronic kidney disease progression, only 22 clinical variables are hand-picked in advance to be used as features.

However, data representation has a big impact on the success of predictive models since learning algorithms, to a large extent, rely on it for capturing what is hidden in the data [Nikravesh et al., 2006]. This is perhaps especially true when learning predictive models from EHRs due to the complex nature of the data. Therefore, insights into how best to represent EHR data for the purpose of learning effective predictive models could yield valuable contributions to the ultimate goal of improving healthcare through analytics [Hripcsak and Albers, 2013]. Although structured EHR data contains predefined variables that, in practice, could be used in a straightforward manner without requiring much feature engineering, many problems remain, in particular with respect to the aforementioned challenges. The
issue of representing structured EHR data, with the aim of handling complexity and exploiting the richness of the data, is currently insufficiently studied. A few studies do address this general issue, e.g., [Wiens, 2014; Singh, 2015], each of which targets, to some extent, one of the described challenges. [Wiens, 2014] tackles temporality by creating multiple binary features for events that occur multiple times within a time period, which, however, results in increasing the dimensionality of the feature space; [Singh, 2015] makes use of the concept hierarchy of clinical codes for dimensionality reduction by merging features (represented by specific codes) from the same family. Both of these studies merely touched upon the issue of representing structured EHR data and did not study it in depth. It is therefore clear that a gap exists with respect to representing structured EHR data that, if filled, may prove useful for learning effective predictive models. This thesis aims to bridge this gap by addressing various technical challenges involved in enabling effective large-scale analysis of EHR data for the purposes of predictive modeling, including:

1. Heterogeneity: EHRs comprise a wide range of heterogeneous types of data, from structured information on e.g. drug administration, diagnoses and laboratory tests, to unstructured information in the form of narrative clinical notes. All of these can be used for learning predictive models; however, the unique nature of each type of data results in unique obstacles with respect to data representation: one cannot assume that identical representations are suitable for diverse types of data. How best to exploit heterogeneous EHR data, individually and in combination, for predictive modeling remains unclear.

2. Concept hierarchies: EHRs typically adopt various encoding systems, such as the International Classification of Diseases\(^1\) (ICD) to encode diagnoses and the Anatomical Therapeutic Chemical\(^2\) (ATC) classification system to encode drugs. These encoding systems contain a hierarchical structure that encodes the concepts, in the form of diseases or drugs, with different levels of specificity, ranging from organ systems to specific local diseases in ICD and from anatomical groups to specific drugs in ATC. The concept hierarchies contain potentially valuable information that may contribute to improving the effectiveness of predictive models. However, it remains

---

\(^1\)Information available at [http://www.who.int/classifications/icd/en/](http://www.who.int/classifications/icd/en/)

\(^2\)Information available at [https://www.whocc.no/atc/structure_and_principles/](https://www.whocc.no/atc/structure_and_principles/)
unclear how the information embedded in the concept hierarchies of clinical codes can be effectively utilized for predictive modeling.

3. Temporality: Clinical events are time-stamped in EHRs. The order of different events, as well as the time distance between them, often carries information that is crucial for analyzing patient trajectories. As a result, it may be advantageous to take temporality into account when analyzing EHR data. However, it is not obvious how best to capture temporal information in predictive models. Moreover, the complexity is further exacerbated by the fact that each patient record effectively consists of multiple time series with different lengths and irregular rhythms.

Note that there are a multitude of other challenges involved in using EHR data. For example, data quality, in terms of imprecision and incompleteness, is a common concern for statistical analysis of EHR data since it may potentially result in biased models [Weiskopf and Weng, 2013]. However, this is not within the scope of the thesis. More details concerning challenges related to analyzing EHR data are described in chapter 2.

1.3 RESEARCH QUESTIONS

The overarching research question that this thesis aims to answer is:

*How can effective predictive models be learned from electronic health records?*

Learning predictive models from EHR data can be studied from various perspectives. This thesis focuses on finding solutions that address the challenges described above with respect to data representation. In an attempt to answer the overarching question, it is broken down into the following related research questions (the papers that contribute to answering each research question are listed in parentheses):

- *How can heterogeneous types of data be represented to allow for effective predictive modeling?* (Paper I, II, III)
- *How can the concept hierarchy of clinical codes be leveraged for effective predictive modeling?* (Paper I and IV)
INTRODUCTION

• How can temporality be incorporated for effective predictive modeling?
(Paper V, VI, VII, VIII)

The effectiveness of a predictive model is here taken as a relative concept in contrast to absolute effectiveness. This means that the aim of this thesis is not to claim whether or not a learned predictive model is effective, but rather to conclude if model A is more effective than model B. The effectiveness is evaluated by comparing the predictive performance of different models, which can be measured using various performance metrics (see subsection 4.2.3 for details).

In this thesis, these questions are studied empirically in the context of pharmacovigilance, also known as drug safety surveillance. Pharmacovigilance concerns the collection, detection, assessment, monitoring, and prevention of adverse effects of drugs [WHO, 2002]. Adverse drug events (ADEs) have become a major public health problem by causing human suffering and resulting in huge healthcare costs [Sultana et al., 2013], often unnecessarily so since many ADEs are known to be preventable [Kanjanarat et al., 2003]. Post-marketing drug safety surveillance has emerged as a vital complement to pre-marketing clinical trials, mainly because many ADEs remain undiscovered during clinical trials, which are limited in terms of both time period and sample size. Electronic health records are considered the most promising data source for post-marketing ADE detection [Chazard, 2011]. A large number of studies have been conducted in recent years on the possibility of identifying potential drug safety issues by using EHR data, see e.g. [Trifiro et al., 2009; Harpaz et al., 2010, 2013; Warrer et al., 2012; Li et al., 2014; Iqbal et al., 2015]. However, it is also well acknowledged that ADEs are heavily under-reported [Classen et al., 2011] or misdiagnosed [Falchuk and Falchuk, 2013] in EHRs. The studies included in this thesis were conducted with the use case of identifying patients with an ADE-specific diagnosis. Most of them focus on distinguishing between patients with an ADE-specific diagnosis code and patients with a similar diagnosis that, however, does not indicate that it was drug-induced. Such models could be used for identifying patients who should have been diagnosed with a specific ADE but were not, thereby improving the reporting of ADEs in EHRs.
1.4 CONTRIBUTIONS

The contributions of this thesis mainly fall within the intersection of computer science and healthcare, an area often referred to as health informatics. Methods are developed to facilitate the identification of features from the richly structured and complex EHR data, which is not only of high dimensionality and sparsity but also heterogeneous and embedded with irregular temporality in terms of length and rhythms. The application of these methods to EHR data pushes forward the meaningful secondary use of EHRs by unlocking more possibilities for analyzing and modeling EHR data. At the more practical end, with the help of the predictive models learned in this thesis, under-reporting of ADEs can potentially be mitigated. The developed methods can be applied in clinical decision support systems that assist doctors in assigning ADE-specific diagnosis codes. Doctors’ awareness of ADEs can hence be enhanced, leading to improved patient safety. Below, a summary of how each paper contributes to answering the research questions is presented.

- **Representing Heterogeneous Types of Data** Due to the unique characteristics of each type of data in EHRs, the representation of data from each type is also specific. For structured EHR data, representing drugs and diagnoses can benefit from exploiting the encoding systems (Paper I), and representing clinical measurements and laboratory tests requires the consideration of missing and repeated events (Paper II). Moreover, in specific cases it is beneficial to combine multiple types of data (Paper III).

- **Leveraging the Concept Hierarchy of Clinical Codes** The impact of using the concept hierarchies of clinical codes — ICD-10\(^3\) for diagnoses and ATC for drugs — has been studied from two perspectives: enriching the feature space and decomposing the predictive task. On the one hand, the concept hierarchy of the two coding systems is shown to be useful for constructing a feature space containing clinical codes from different hierarchical levels to improve the predictive performance (Paper I). On the other hand, the concept hierarchy of ICD-10 codes is shown to be beneficial for decomposing the predictive task into a number of sequential steps, serving different application scenarios such as predicting a disease family or a specific disease (Paper IV).

---

\(^3\)The 10th edition of the ICD encoding system.
• **Incorporating Temporal Information** Temporal information is shown to be valuable for learning effective predictive models. Various ways of incorporating temporality have been investigated in this thesis by creating a single-point representation or a multivariate time series representation. In the single-point representation, by assigning temporal weights to clinical events prior to learning the predictive model, the predictive performance is improved compared to ignoring temporality or treating events from different time windows as different features (Paper V). Moving forward in this direction, instead of using pre-assigned temporal weights (various weighting strategies are explored in Paper VI), it is shown to be even more effective to learn the temporal weights while learning the predictive model (Paper VII). In the multivariate time series representation, each clinical event is treated as a time series, which leads to an advanced representation consisting of multivariate time series that take into account both the order of clinical events and their corresponding values. Methods are developed to build predictive models using such a representation, which are useful not only in representing longitudinal EHR data but also for general time series classification problems (Paper VIII). All the developed methods succeed in capturing the temporal information and are compatible with any standard learning algorithm.

### 1.5 DISPOSITION

The remainder of the thesis is organized as follows. An extended background is presented in chapters 2 and 3, where chapter 2 provides a general introduction to electronic health records with a focus on meaningful secondary use, challenges of analyzing EHR data and a brief description of the encoding systems used in EHRs; chapter 3 introduces predictive modeling and other basic concepts in machine learning, in particular supervised machine learning, data representation and related research on representing EHR data for building predictive models. In chapter 4, the methodology used in the thesis is described, including research strategy, experimental setups, evaluation and ethical issues. The developed methods and models are evaluated empirically, and the application area, data source, a summary of study designs and main findings of the included papers are described in chapter 5. Finally, concluding remarks are made in chapter 6, where answers to the
research questions are presented, followed by a general discussion, as well as conclusions and suggestions for future research.
CHAPTER 2

ELECTRONIC HEALTH RECORDS

Electronic health records (EHRs), replacing traditional paper-based patient records, collect and store patient data systematically over time in a digital format. Thanks to initiatives such as the Health Information Technology for Economic and Clinical Health Act (HITECH) [Blumenthal, 2010; Blumenthal and Tavenner, 2010] in the United States and the Innovative Medicines Initiative (IMI) [Hunter, 2008] in the European Union, EHRs are increasingly being adopted worldwide. In smaller European countries like Sweden, a high level of paperless patient data sharing, storage and decision support is being reached well ahead of larger countries. The use of EHRs enhances the quality of healthcare, simplifies communication across different healthcare settings through paperless data sharing, and enables the reuse of patient data on a large scale. In this chapter, the characteristics of EHR data are first described, followed by the opportunities and challenges of facilitating meaningful secondary use of EHRs through large-scale analysis.
2.1 DATA CHARACTERISTICS

A wide range of types of data are often included in an EHR system, including demographic data, diagnoses, drug administrations, allergies and immunization status, clinical measurements, laboratory tests, radiology images, billing information, and clinical notes. Since EHRs are primarily used for administrative purposes, demographic patient information, such as age and gender, is typically available for most patients. Some other types of data are generated by physicians placing electronic requests for services from other medical departments. In EHRs, physicians place requests for ancillary hospital activities that are provided by laboratories, pharmacies and radiological departments through computerized physician order entry (CPOE) systems. These are computer-based systems that share the feature of automating the ordering process, which not only provides convenience for physicians but also diminishes the risk of human mistakes like miscommunication. For example, adopting CPOE systems for electronic prescriptions greatly reduces the risk of medication errors [Ammenwerth et al., 2008]. The different types of data in EHRs have their own characteristics, e.g. medications and diagnoses are often encoded with standard coding systems while clinical notes are in free-text. Below, the characteristics of several basic types of data are described.

Medications

To make sure that the CPOE systems work smoothly, standard encoding systems are needed to avoid terminological mismatching. This is especially important for prescribing medications electronically since there are many brands of drugs that share exactly the same active substances. The Anatomical Therapeutic Chemical (ATC) classification system is used in most EHRs [WHO, 1993]. It divides the active substances into different groups in a hierarchical fashion, described below.

- 1st level – indicated by a single Latin letter. Medications are divided into 14 anatomical groups. For example, C: cardiovascular system.
- 2nd level – indicated by two digits. Each anatomical group is divided into different therapeutic subgroups. For example, C10: lipid modifying agents.
• 3rd level – indicated by a single Latin letter. Each therapeutic group is divided into different pharmacological subgroups. For example, \textit{C10A}: lipid modifying agents, plain.

• 4th level – indicated by a single Latin letter. Each pharmacological group is divided into different chemical subgroups. For example, \textit{C10AA}: HMG CoA reductase inhibitors.

• 5th level – indicated by two digits. Each chemical group is further divided into different chemical substances. For example, \textit{C10AA01}: Simvastatin.

Diagnoses

The International Statistical Classification of Diseases and Related Health Problems (ICD) system is an internationally recognized classification that helps physicians, patients and policy makers to navigate and communicate between healthcare systems, as well as for billing purposes. The ICD system is revised regularly to meet expanding disease classification needs [Maxim Topaz MA and Shafran-Topaz, 2013]. The current widely used version is the 10th edition, ICD-10 [Steindel, 2010]. Each specific disease is encoded by a four-level hierarchical code in the ICD-10 system, each of which is described as follows.

• 1st level – indicated by a single Latin letter. Diseases are divided into 22 chapters according to organ system or etiology. For example, \textit{F}: mental and behavioral disorders.

• 2nd level – indicated by a single digit. Each chapter is divided into blocks of disease families. For example, \textit{F2}: schizophrenia, schizotypal and delusional disorders.

• 3rd level – indicated by a single digit. Each block is divided into subgroups containing specific diseases. For example, \textit{F25}: schizoaffective disorders.

• 4th level – indicated by a single digit after a decimal point. Diseases from the same subgroup are further divided according to different types or stages of a disease. For example, \textit{F25.1}: schizoaffective disorder, depressive type.
Clinical measurements and laboratory tests

Clinical measurements include basic measurements that are carried out during patient visits, such as body temperature, body weight and blood pressure. Laboratory tests are the more advanced measurements that are requested by physicians and carried out by different laboratories, such as chemistry labs and pharmacist labs. Standard codes are also used to represent each measurement and test, such as the Logical Observation Identifiers Names and Codes (LOINC). Depending on the type of measurement or test, the reported results can be in the form of both numerical and categorical values.

Clinical notes

Since the above types of data are reported by physicians through standard interfaces in EHRs, information that is not required but yet important is then reported in free-text such as admission notes, treatment plans and patient summaries. Clinical notes are perhaps the most abundant type of data in EHRs, but they are also arguably the most difficult to analyze computationally. Free text allows the flexibility to report any information that is important, but it also introduces challenges such as handling spelling errors, abbreviations and acronyms, and incomplete sentences.

The constant availability of these basic types of data in most EHRs makes them a popular source of phenotype information in population-level research using EHRs [Jensen et al., 2012]. Other information, such as radiology images, is also important for supporting the diagnosis procedure. The various types of data complement each other to present a more complete picture of patient trajectories. The availability of large amounts of clinical data has triggered increasingly more researchers to explore ways of making meaningful secondary use of EHRs.

2.2 MEANINGFUL SECONDARY USE

The widespread employment of EHRs generates an unprecedented amount of clinical data during routine clinical care, which can be accessed by researchers under agreed conditions including ethical permissions. This accessibility triggers increasing interest in the secondary use of EHR data, beyond using EHRs merely for archiving and communication purposes. The systematic collection of data
in EHRs provides the possibility for improving healthcare through enhanced error detection, improved adherence and reduced costs. The obvious benefits of using EHRs include quality measurements, public health surveillance and patient access to their health status [Safran et al., 2007]. Besides these, the biggest promise of using EHRs lies in facilitating large-scale analysis, creating a clinical data repository and hence enabling more efficient retrospective clinical research [Hripcsak and Albers, 2013].

As a result of the digitization of clinical documentation, enormous amounts of clinical data become processable by computers. This, in turn, means that data from millions of patients across a region or even countries has the potential to be used for research on a large scale. By having access to EHR systems, clinical and biomedical research can more readily be carried out at the population level [Hersh, 2007]. With the advances of research in the data science\(^1\) area, the robust capacity of data storage, management and analysis provides opportunities to make use of large amounts of clinical data in a more effective and efficient manner. What we lack in knowledge can now be compensated, to some degree, by looking at abundant data. Moreover, large-scale analysis goes beyond the possibility of analyzing large amounts of data; it also means the potential of analyzing different kinds of data, in contrast to the typical numerical or categorical structured data that can be analyzed by standard statistical methods. For example, with the help of natural language processing (NLP) techniques [Manning and Schütze, 1999], clinical notes in the form of free-text can be read and processed by computers and eventually analyzed to extract useful knowledge; deep learning techniques [LeCun et al., 2015] largely improve the capacity of computers to process medical images to detect abnormalities.

EHR data comes in the form of longitudinal observations from various medical departments. Observations are often recorded for millions of patients, spanning many years. This level of data collection is impossible to achieve by clinical trials, in which a small sample of patients who are willing to participate in the research are observed for a limited time period. Because prospective collection of clinical data is notoriously expensive and time-consuming, extracting relevant data from EHRs allows researchers to develop a clinical data repository containing extensive records for large numbers of patients.

---

\(^1\)Data science is an area that concerns methods and techniques for collecting, organizing and analyzing data with the purpose of extracting new knowledge.
With the clinical data repositories, retrospective clinical research can be carried out more efficiently and with potentially improved outcomes, especially in areas of public health like epidemiology and pharmacovigilance. Retrospective analysis of EHR data can help improve statistics on disease incidence and enhance our understanding of disease development to a large extent at a relatively low cost. In the long term, analyzing EHR data retrospectively also opens up the possibility of combining phenotypic data with genetic data for novel discovery of associations between diseases and genetic or environmental factors [Denny, 2012; Pathak et al., 2013].

2.3 CHALLENGES AND PITFALLS

Meaningful secondary use of EHRs typically requires high-quality data, which are often lacking in EHRs [Weiskopf and Weng, 2013; Bayley et al., 2013; Paxton et al., 2013]. The challenges of analyzing EHR data in clinical and biomedical research primarily concern completeness, correctness and complexity.

- **Completeness**: This means that everything that is known about a patient is not always completely presented in the EHR database. Even with a standard user interface collecting patient data systematically, EHRs still suffer from under-reporting. There are various reasons for this, e.g., (1) lack of meticulous recording of patient information in EHRs; (2) doctors failing to identify specific conditions such as adverse drug events [Hazell and Shakir, 2006]; (3) poor communication between doctors when more than one doctor is involved in one patient record; and (4) difficulties in integrating the different EHR systems that are used by different hospitals, which results in a chunk of medical history being lost when patients switching hospitals.

- **Correctness**: This refers to whether certain information reported in EHRs about a patient is correct. Imprecise information often exists in the shape of information not being specific enough, instead of being completely wrong, which is quite rare. For example, a patient record may contain a diagnosis of a general skin disorder while the real underlying condition is a drug-induced skin disorder. According to the Swedish National Board of Health and Welfare, in a review of 4200 health records, about 20% of primary diagnosis
codes assigned in EHRs were found to have major errors [Socialstyrelsen, 2006].

- **Complexity**: This refers mainly to the complexity of the data in EHRs. The existence of heterogeneous types of data in EHRs highlight the difficulties in analyzing them. The difficulties are present not only in combining the heterogeneous types of data, but also in analyzing them separately. To process the clinical notes in free text, natural language processing is needed, and the tasks are complicated by the low quality of texts containing massive amounts of medical abbreviations, typographical errors and incomplete sentences. The clinical events that are reported in EHRs, including diagnoses, medications and laboratory tests, are often temporal in nature. Often more than one type of clinical event is reported in each patient record, and each clinical event is reported more than once depending on how often a patient visits the hospital. Except for some regularly hospitalized patients, the time gap between two hospital visits often varies. Therefore, each patient record consists, in effect, of multiple time series with different lengths and irregular intervals, which makes it particularly difficult to capture the temporal information. When preparing EHR data for analysis, some other technical challenges are also often encountered, including high dimensionality due to the large number of unique clinical events; high sparsity due to each clinical variable being shared by only a small group of patients; highly skewed data due to very few health records containing the target outcome of interest compared to those that do not.
CHAPTER 3

PREDICTIVE MODELING

Machine learning is playing an increasingly crucial role for large-scale data analysis. Predictive modeling uses machine learning algorithms to learn from historical data and make predictions on new data. This provides great opportunities for knowledge discovery and decision support. This chapter begins with a brief introduction to machine learning. Among the different types of learning that exist, the supervised learning that is used in this thesis is introduced. The importance of data representation is then highlighted as this is the cornerstone of predictive modeling and is related to the primary contribution of this thesis. In the end of this chapter, previous research on predictive modeling using EHR data is presented, in particular with respect to data representation methods that address the challenges of analyzing EHR data.
Chapter 3.

3.1 MACHINE LEARNING

Machine learning is an area within artificial intelligence that aims to answer the question: “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” [Mitchell, 2006]. Learning algorithms are used for predictive modeling to learn relationships from historical data in order to make predictions\(^1\) on new data. A learning algorithm typically takes a number of examples or instances – the things that are to be, e.g., classified or clustered – as input, where each example is characterized by its values for a predefined set of features or attributes. For example, in a dataset with examples corresponding to cars, the features might be brand, color, horse power, etc. The output is the knowledge that is learned from the input data, which is often referred to as training data, and can be represented in various ways, such as by decision trees or classification rules [Witten and Frank, 2005]. The quality of a learning algorithm is often determined by two aspects: efficiency and effectiveness. The former concerns the computational cost that is commonly reflected by the memory consumed and the time taken to complete the learning procedure. The latter can be reflected by the performance of the model built by the learning algorithm; model performance is usually measured on a set of data that was not used for training, often referred to as test data.

Based on the availability of input data, machine learning can be categorized into three types: supervised learning, unsupervised learning and semi-supervised learning. In supervised learning, the input data is labeled and the algorithm learns the relationships that map input data to the labels; in unsupervised learning, no labels are given and the algorithm learns structures in the input data; in semi-supervised learning, only a part of the input data is labeled and the algorithm makes use of both labeled and unlabeled data to learn labels for the unlabeled part/relationships between input data and the labels. Supervised machine learning is primarily introduced here since this is the only type of learning setting that is covered in the thesis. However, note that unsupervised and semi-supervised learning are also very useful for analyzing EHR data. In particular, they are often applied for analyzing clinical notes, where there are no or a very small amount of annotated notes available, see e.g. [Wang et al., 2012; Garla et al., 2013; Henriksson, 2015a].

\(^1\)To many, prediction refers to future events; however, in machine learning, this is a more general concept that applies to any type of unknown event, regardless of when it occurred.
Supervised machine learning

In supervised learning, given \( N \) training examples \( \{x_1, x_2, \ldots, x_N\} \in X \), each of them has a label representing desired output \( \{y_1, y_2, \ldots, y_N\} \in Y \). A perfect mapping from the input space \( X \) to the output space \( Y \) is the target function \( f \), which is unknown. The learning algorithm hence seeks a function \( g : X \rightarrow Y \) that approximates \( f \) using the training data. There are typically many possible functions that can realize such approximations, which together comprise the hypothesis space \( G, g \in G \). The choice of the final hypothesis \( g \) from \( G \) is by

\[
g_\theta(x) = \arg\max_y s_\theta(x, y),
\]

(3.1)

where \( \theta \in \Theta \) defines the model parameters and \( s \) is a scoring function, \( s : X \times Y \times \Theta \rightarrow \mathbb{R} \).

Supervised learning tasks are divided into two categories based on the data type of the label: when the label is categorical, it is a classification task; when the label is numerical, it is a regression task. For example, given a group of patients who are diagnosed with different diseases together with their laboratory tests results, a classification task can be to predict which disease a new patient should most probably be diagnosed with; a regression task can be to predict the score of a blood test on a new patient. In general, a supervised learning task aims to minimize the expected loss, such as prediction error or squared error. For classification, which is the focus of this thesis, when the label consists of only two categories, it is a binary classification problem. The examples that are labeled as belonging to the category of interest are often referred to as positive examples, and the rest as negative examples. When the label consists of more than two categories, it becomes a multiclass classification problem. A multiclass problem can be transformed into a set of binary problems by employing the one versus all strategy [Rifkin and Klautau, 2004], and in each of the binary problems, the examples belonging to a particular class are the positive examples and the examples of all other classes are the negative examples.

Applying supervised machine learning to electronic health records

Working with real world data typically entails working with data that cannot be used directly by a learning algorithm since it has not been collected for that primary purpose. There are often established steps, such as the CRISP-DM\(^2\) model

\(^2\)Cross Industry Standard Process for Data Mining
that can be followed when analyzing the data to solve a real world problem. In the case of applying supervised machine learning to EHR data, there are a few specific questions that first need to be answered.

- **What is the label?** Based on the targeted outcome of the predictive task, one can decide what the label is, which will help in extracting training data from the EHR database. For example, when the target outcome is related to a specific disease, the corresponding diagnosis code representing this disease can be used as the label and the health records that contain this code can be extracted as part of the training data.

- **What is a training example?** When the label has been chosen, one needs to decide what constitutes an example related to the label. For example, in EHR data, a typical training example can be a patient or a care episode. The choice is often made according to the predictive task, e.g., if the model is to predict whether a patient will develop the disease or to predict whether the diagnosis code indicating the disease should be assigned in a care episode. It is sometimes not so obvious what a care episode is. Lacking a standardized definition, different studies employ their own definitions of a care episode, such as a health record that has clinical events reported continuously without a time gap longer than 24 hours [Henriksson et al., 2015a,b, 2016].

- **What type of data to use?** As introduced before, EHRs contain heterogeneous types of data, which can be used separately or in combination. The choice of type(s) of data will largely influence the required techniques for pre-processing the data and, perhaps also, the choice of learning algorithm. For example, if clinical notes in free-text are to be used, natural language processing methods are then needed to pre-process the data.

- **How much of the patient history to include?** Clinical data is collected over time in EHRs. Some patients have a medical history of many years stored in the EHR system. It is often not realistic or necessary to use the whole patient history; instead a subset of it can be used, depending on the predictive task. Sometimes it can be an absolute time period, e.g., from May to August of each year, or a relative time period, e.g., three months before the diagnosis of a specific disease for each patient.
After having answered these questions, one is only one step away from having the data that can be used to build predictive models: what are the features? The creation of features from raw data is related to the activity called data representation.

3.2 DATA REPRESENTATION

When beginning to work on a predictive modeling problem, one important step is to represent the raw data in a form that can be recognized by the learning algorithm, which is also known as feature construction. A representation of data is a set of features that are constructed from the raw data, which often belong to different data types such as numerical values, categorical values, histograms, or time series.

The success of a machine learning algorithm depends, to a large extent, on representing the data in such a way that useful information hidden in the data can be exploited by the algorithm [Bengio et al., 2013]. It was estimated that data preparation accounts for around 60% of the effort in a data mining task [Cabena et al., 1998]. Typically, a learning algorithm requires input that is computationally convenient to process. However, in many real life situations, such as speech recognition, natural language processing and image processing, the data is often not in a form that is suitable to feed to a machine learning algorithm. Therefore, transforming the raw data to construct relevant and informative features is key to effective predictive modeling. Feature construction can take advantage of human intelligence (e.g., prior knowledge on the data or task) and/or artificial intelligence to assist learning algorithms to extract and organize the discriminative information from the data. It can be done either by manually creating features from raw data or automatically discovering useful features from raw data. The former is straightforward but relies heavily on human knowledge and experience, and sometimes limits the discovery of underlying explanatory factors hidden in the data. The latter, increasingly more popular nowadays, typically deploys machine learning algorithms to extract useful representations of features from raw data.

Learning representations corresponds to the ultimate goal of making machine learning less dependent on manual feature construction in the progression toward artificial intelligence. This concept is often related to deep learning methods [Bengio et al., 2012, 2013], where multiple layers of nonlinear transformations are
composed to yield more abstract representations. Besides deep learning, there are also various other ways of learning representations. Principal component analysis (PCA) [Wold et al., 1987] is one of the earliest feature extraction methods. It is often used for dimensionality reduction by transforming the original features into principal components and selecting only a subset of those components that have the highest variance, such that as much variation in the data as possible is accounted for. Independent component analysis (ICA) [Hyvärinen and Oja, 2000] is a technique for learning a data representation by using a weighted sum of independent non-Gaussian components. Clustering algorithms, such as k-means clustering [Hartigan and Wong, 1979], can be used to learn data representations by grouping inputs into clusters and then using the centroids of these clusters to produce features. When the raw data contains complex structures such as temporal or spatial information, a representation of such data can be learned in a two-step fashion: first, a naïve representation is created, e.g., binary features for each time point or location; then, the hidden relationships among the naïve features are learned to create more advanced features by using methods like variable importance analysis from the random forest algorithm [Breiman, 2001]. The second step is sometimes related to feature weighting [Wettschereck et al., 1997].

What makes a representation good? It is a straightforward question to ask, yet the answer to which is far from obvious. One may, for instance, argue from the perspectives of interpretability, complexity, efficiency and effectiveness.

- **Interpretability** refers to what extent a representation can be understood and be made sense out of in the context of the predictive task, as well as if a representation incorporates the prior knowledge one possesses of the data and the task. Sometimes we talk about the interpretability of a predictive model, which to a large extent requires an interpretable data representation.

- The **complexity** of a representation mainly refers to how well it captures the underlying structure of the raw data in terms of, e.g., temporal and spatial coherence. Temporality and spatiality are difficult for a representation to capture without the expense of creating millions of binary features. It often requires the constructed features to take on a more complex shape, such as histograms or time series, rather than single values.
Efficiency mainly refers to how manageable a representation is for an algorithm from the point of view of computational cost\(^3\). It is often related to the dimensionality, which is the number of features identified from the raw data. A representation with many features can cause the so-called curse of dimensionality [Bishop, 2007], meaning that the input to an algorithm is too large to process and is also likely to be redundant. Sparsity is often related to high-dimensional data. For any given example in the data, only a small fraction of the features are relevant, meaning that a large portion of the features are zero. This results in most of the constructed features being insensitive to small variations.

The effectiveness of a representation concerns the informativeness of the constructed features. It can be evaluated by calculating feature importance and predictive performance scores of a predictive model built using the representation.

When creating a data representation, there are often trade-offs between the above perspectives that need to be considered. For example, a representation that is highly effective might come at the expense of a loss in interpretability. This is especially true in tasks such as image recognition, where a representation learned through deep learning often leads to a very effective predictive model; however, such a representation is often very abstract and hence difficult to interpret. In some cases, interpretability is sacrificed by instead aiming for a more efficient representation. To improve efficiency, dimensionality reduction techniques are often applied. When data is projected from a higher dimension to a lower one, the link between the two dimensions is not always traceable, meaning that there is no way to explain the prediction by features from the original representation. Random indexing is one such technique, which is often used in text mining for word representation [Sahlgren, 2005].

Unfortunately, it is difficult to find a standard merit to assess the quality of a representation. Most commonly, it is done by quantitatively evaluating its effectiveness. In this thesis, this approach is used when comparing different data representations.

\(^3\)Even though there is a cost involved in deriving the features, e.g., through PCA, in the data representation procedure, here the concern is mainly about the efficiency for a learning algorithm to use the constructed features.
3.3 RELATED RESEARCH

Due to the requirement of an ethical permission to access EHR systems, research on predictive modeling using EHR data is nascent. Most studies that analyze EHR data using machine learning focus on applying certain learning algorithms to predict a specific disease, where the data used often contains predefined features; here, however, consideration is primarily given to studies investigating data representation methods when extracting datasets from EHRs that can be exploited by machine learning algorithms. The selected studies are presented in relation to the research questions posed in this thesis.

Many studies use clinical codes for representing drugs and diagnoses. Diagnosis codes have been used as features to detect epidemics [Tsui et al., 2001], diabetes [Krishnan et al., 2013], heart failure [Sun et al., 2012], hospital-associated infections [Wiens et al., 2012] and drug-induced skin disorder [Karlsson et al., 2013]. In these studies, each diagnosis is represented as a binary feature indicating its existence or absence in a health record, where either the most specific level or a more general level of the code is used. The hierarchical structure of the diagnosis codes has also been leveraged by combining the most specific level and a more general level to create a new feature set [Singh and Guttag, 2013]. A similar aggregation of codes using a coding system’s hierarchical structure has also been applied on drugs [Chazard et al., 2011]. However, no evidence that supports the choice of these particular levels has been presented, e.g., demonstrating their superiority over other levels in terms of model performance. When representing clinical measurements or laboratory tests, studies differ in the degree to which temporality is taken into account. Some solutions completely ignore it by roughly aggregating a clinical measurement with a set of different values collected over time into a binary one, such as hyperkalemia (1 = too high potassium value; 0 = otherwise) [Chazard et al., 2011].

To handle temporal data in terms of data representation, some studies divide the whole time period into a number of time windows. For example, in a study to detect signals of the side-effects of drugs, the temporal problem is transformed into a cross-section problem by defining a hazard period, effect period and reference period after a drug is prescribed; temporal association rules are then created in each period [Jin et al., 2010]. In other studies, a stacked-temporal strategy is developed, which divides the patient history into a number of time windows and then creates features from each window that are further concatenated to form...
the final feature set [Wiens, 2014; Singh et al., 2015]. There, the authors also propose a multitask-temporal strategy, where a predictive model is created using features from each time window and then aggregating the outcomes from each model. However, in this case, the problem is segmented into multiple, parallel and independent tasks, where each task corresponds to a different time period, which fails to consider the full picture of the whole time period. In some application areas, such multitask strategies can be problematic. For instance, to detect chronic diseases using clinical data from patient history, the target disease might result from co-occurrences of clinical events from different time windows; however, by creating multitasks from a sequence of small time windows, such co-occurrences are ignored.

There are also studies that look at the whole time period and treat a set of temporal data as time series, where the problem boils down to generating features that represent these time series [Patel et al., 2008; Batal et al., 2012, 2013; Moskovitch and Shahar, 2015; Moskovitch et al., 2015; Liu et al., 2015]. Some of these [Patel et al., 2008; Batal et al., 2012, 2013] build on temporal abstraction [Shahar, 1997] or temporal logic [Allen, 1984] to define patterns that can describe temporal relationships among multiple time series. Such methods result in temporal patterns, such as “the occurrence of clinical event A precedes a decrease in clinical event B”; these are then analyzed to find the most informative patterns for the classification task. Other studies [Moskovitch and Shahar, 2015; Moskovitch et al., 2015] transform the time-stamped data points into symbolic time intervals and then discover frequent time-interval-related patterns that are used to induce a classifier. Yet other methods [Liu et al., 2015] involve creating a graph-based framework, where the temporal relations are maintained as temporal graphs.

One popular representation of time series is to transform a time series into a symbolic sequence, typically as a string consisting of symbols. This strategy allows various ways of creating features using the generated strings. For example, a bag-of-words representation can be created by using a sliding window to transform a time series into symbols and extracting all unique sequences to form a pattern vocabulary [Ordóñez et al., 2011; Lin et al., 2012]. An alternative is to adopt so-called shapelet-based methods, where a shapelet is a subset of a time

---

4The bag-of-words representation is a simplifying representation used in natural language processing, where a text (such as a sentence or a document) is represented as a bag of the words it contains, disregarding grammar and even word order but keeping multiplicity. It has been generalized as a representation method for data other than language. Here, it means that a multiset of features are used regardless of their order or dependency.
series. The single-scan shapelet transformation algorithm looks for the best $k$ shapelets from a set of univariate time series, and then transforms the original time series data into a dataset with $k$ features, each of which is the distance between the series and one shapelet [Hills et al., 2014]. The shapelet transformation algorithm together with a C4.5 decision tree demonstrated competitive accuracy against the original shapelet tree by Ye and Keogh [Ye and Keogh, 2009]. In the setting of their study, each object is only described by a single time series; with a reasonably small $k$, the transformed dataset still has a manageable dimensionality. However, in the situation where an object is described by a large set of time series, the dimensionality of the transformed dataset will be expanded $k$ times. The effect of the curse of dimensionality will hence be amplified. In the case of multivariate time series, the generated candidate shapelet is also multivariate and has the same length in each dimension [Ghalwash and Obradovic, 2012].
CHAPTER 4

METHODOLOGY

Research is defined, in the Oxford Dictionary of English, as “the systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions”. A systematic investigation implies the use of research methods that provide guidance in the search for facts and conclusions. In this chapter, the research strategy that is followed in the thesis is described, including a discussion of the underlying philosophical assumptions and questions concerning reproducibility, reliability and generalizability. In order to answer the research questions, empirical investigations are carried out in which the proposed solutions are evaluated. The key components of the experimental setups used for evaluation – learning algorithms, performance metrics and hypothesis testing – are also presented in this chapter. Finally, relevant ethical issues are discussed.
4.1 RESEARCH STRATEGY

The research paradigm that underpins the methodology employed in this thesis is positivism. The positivist research paradigm subscribes to an empiricist view of epistemology and hence emphasizes obtaining knowledge through observation and experimentation [Cohen et al., 2013]. Research within this paradigm relies on the collection of verifiable empirical evidence in support of theories or hypotheses, which is in accordance with the deductive nature of research in machine learning, where a hypothesis often is developed based on existing theory and tested empirically or by means of mathematical proof. In this thesis, hypotheses are tested empirically through experiments. In experimental research, observations are made in a controlled environment in order to determine the effect of one or more variables, which are manipulated, on a dependent variable. Here, the observations are in the form of either predictions for a single dataset or in the form of predictive performance estimates over multiple datasets. The input variables under investigation are data transformation strategies and the dependent variable is in the form of predictive performance. Predictive performance is typically measured using performance metrics in the form of numerical scores that are obtained by comparing predictions to ground truth, e.g. in the form of known outcomes. In order to verify that the made observations are statistically significant and not due to chance, the hypotheses are tested using statistical inference. Since the research questions concern evaluating the relative effectiveness of predictive models, the hypotheses are generally in the following form: model A is more effective than model B, where the null hypothesis is that there is no difference between the two models. The method for the hypothesis testing is chosen with consideration of two main factors: (1) the number of datasets from which observations are generated; (2) the number of models involved in the comparison.

Although the research is primarily deductive and causal, it is also, in some sense, exploratory: it is not always obvious from the outset which precise problems are to be solved, especially when facing an emerging area – like large-scale analysis of EHR data – where there is the potential for addressing a multitude of challenges. On the other hand, when tracking down the solutions for a specific challenge, it is often only after one problem has been solved that another emerges. This thesis starts out by exploring the problem area on a general level by covering different aspects in parallel. Based on the conclusions drawn on that level, the exploration is then deepened in order to understand the made observations better.
More specifically, different types of EHR data are studied separately in Paper I and II. A series of follow-up studies – Paper V to VIII – are then conducted based on the observations from Paper III, which builds upon the first two papers. Paper IV takes on the problem from a slightly different perspective by instead manipulating the representation of the class label in order to decompose the predictive task.

When developing a research strategy, it is important to consider (i) reproducibility – the extent to which the empirical evidence can be reproduced, (ii) reliability – the extent to which the empirical evidence reflects the true state of affairs within its own setting, and (iii) generalizability – the extent to which the research findings can be applied to settings other than those in which they were originally tested. Ideally, in order to reproduce the empirical evidence of a study, a detailed description of the experimental setup and the data used in the experiments should both be provided. However, due to the sensitivity of EHR data, it is often not possible to make the employed datasets publicly available. If the research focuses primarily on method development, rather than a specific output highly related to the used data, it is common to provide only a description of the experimental setup. In this thesis, all conducted experiments follow strictly designed experimental setups, which endows the research with reproducibility. To ensure reliability, the size and number of datasets are important factors. Although the size of the training data is likely to affect the performance of the generated predictive models – and, by extension, the effectiveness of the utilized data representation – obtaining reliable performance estimates primarily depends on the size of the test data. Moreover, it is important to evaluate predictive models on independent test data: if the same data is used both to build and evaluate the model, the generated observations are likely to be over optimistic with respect to model performance and consequently the trained model will most likely perform worse on unseen data. To avoid over-estimated performance estimates, the data can be divided into training and test sets. Training data is used to train a predictive model, while test data is used for generating observations in the form of predictions. However, there is a trade-off between the size of training and test data: more data in the training set typically enhances the performance of the predictive model, but more data in the test set reduces the risk of unreliable performance estimates. An alternative is to conduct k-fold cross validation. It divides the data into k folds and uses one of the k folds as test data and the remaining k – 1 folds as training data. To reduce variability, the cross validation procedure includes k rounds until every fold has been used as test data once, and the final outcome is averaged over the rounds. The generalizability, to a large extent, is entwined with reliability. By
evaluating the learnt predictive models on independent test data, the finding that one model outperforms another based on the made observations in an experiment is more reliable. This, in turn, increases our confidence that the same model will continue to outperform the other model on data sampled from the same underlying distribution. However, any observed findings are always conditioned on the used dataset and the task the models are applied to. Therefore, it is not reasonable to assume that an optimal model can be learned that will invariably outperform the others in a different domain with data drawn from a different distribution.

4.2 EVALUATION

In this thesis, the developed methods are evaluated empirically and quantitatively by comparing the predictive performance of the generated models in a controlled experimental setting.

4.2.1 EXPERIMENTAL SETUPS

The experimental setups employed in the included papers can be summarized according to the following steps: (1) data extraction, (2) data representation, (3) classification, (4) predictive performance evaluation and (5) statistical significance testing, as illustrated in Figure 4.1. Note that the majority of this thesis focuses on “transformation strategies” by proposing various strategies for representing EHR data, while some parts of the thesis are closely related to “classification” (e.g. Paper IV).

Data extraction Raw data is extracted from an EHR database according to the selection of types of data. In this thesis, structured EHR data consisting of diagnoses, drugs and clinical measurements is extracted and used for learning predictive models.

Data representation The extracted raw data is transformed into datasets that can be directly employed by standard learning algorithms, which typically requires that the data is in a tabular format with rows comprising training and test examples and columns comprising features. Different strategies are proposed in this thesis for representing the raw data. The transformation strategies are evaluated by
comparing the predictive performance of the predictive models that are built from datasets that have been derived using the different strategies.

Classification Predictive models are built using the underlying learning algorithms on the datasets that have been derived using the different strategies. The choice of underlying learning algorithms is described in detail in subsection 4.2.2.

Predictive performance evaluation Based on the predictions produced by the predictive models, predictive performance scores are calculated with respect to the chosen performance metrics. The chosen performance metrics are listed in subsection 4.2.3. To ensure that models are evaluated on independent data and to
avoid unreliable performance estimates due to chance, 10-fold cross-validation or 5-fold cross-validation with 2 iterations is employed when evaluating the models in this thesis.

**Statistical significance testing** The statistical significance of the observed differences between predictive models that are built on datasets with different representations is tested using different hypothesis testing methods according to the number of datasets and models involved. See subsection 4.2.4 for the choice of hypothesis testing methods.

### 4.2.2 LEARNING ALGORITHMS

The primary underlying classifier used in this thesis for generating predictive models is the *random forest* learning algorithm [Breiman, 2001]. Random forest is an ensemble learning algorithm that generates a set of decision trees, where each tree is built using a bootstrapped sample of the original training examples and each node in the tree is obtained from considering only a randomly selected subset of all available features. In this way, the diversity among trees is established, which is key in the random forest algorithm. Each tree is generated independently and the forest eventually makes the final decision on which class label to assign through majority voting by all of the trees. The ensemble error rate depends not only on the average error of individual base models, but also on their diversity [Krogh et al., 1995]. In other words, given that each tree performs better than random guessing, when the number of trees in the forest increases, the probability that a majority of trees makes an error decreases. Support for this assumption can be traced back to Condorcet’s jury theorem [Condorcet, 1785], which states that the error of the majority of a jury decreases as the number of jury members increases. Therefore, two important factors for the random forest algorithm are individual tree quality and diversity. The estimation of individual tree quality is straightforward by comparing the predicted label and the true label of the samples involved in a single tree. For diversity, the estimation is more complicated. In a regression task, i.e., when the predicted outcome is numeric, the ensemble error $E$ is directly related to the average error of the individual trees $\bar{A}$ and their diversity.
\( \bar{D} \), where diversity is estimated by the average deviation of each single prediction from the ensemble prediction (Equation 4.1).

\[
E = \bar{A} - \bar{D}
\]  

(4.1)

As a result, the ensemble error can never be higher than the average error of the individual trees, and the greater the diversity is, the lower the ensemble error will be. Since two very accurate trees will typically agree on most, if not all, of the predictions, it is not always obvious that one can use the above equation to search for an optimal ensemble. However, this framework does not directly apply to classification, i.e., when the predicted outcome is a label. Alternative ways of estimating diversity can be found in [Kuncheva and Whitaker, 2003].

With its reputation of a relatively low computation cost and robust predictive performance, the random forest algorithm has become one of the most popular learning algorithms [Caruana et al., 2008], especially when working with high-dimensional biomedical data, e.g., microarray data [Díaz-Uriarte and De Andres, 2006]. In all of the studies included in this thesis, predictive models were generated using the random forest algorithm with 500 trees, where \( \sqrt{N} \) features are randomly selected to determine each node in each tree, where \( N \) is the total number of features.

The random forest algorithm provides the possibility of obtaining estimates of variable importance. There are different ways to estimate variable importance (see e.g., [Breiman, 2001]). In this thesis, Gini importance [Strobl et al., 2007] is chosen as the variable importance metric. It is generated by calculating the Gini impurity index, a computationally efficient approximation of the entropy that measures to what extent a potential split is able to separate the samples of the classes during training. A high Gini importance score indicates that a variable plays a greater role in splitting the data into the defined classes, while a Gini importance of zero means that a variable is considered useless or is never selected to build any tree in the forest. One advantage of using Gini importance is that it is relatively cheap to obtain; it has also been successfully used to conduct feature selection on high-dimensional data, such as microarrays [Li et al., 2005; Díaz-Uriarte and De Andres, 2006], time series [Shen et al., 2007; Menze et al., 2007] and spectra [Geurts et al., 2005; Granitto et al., 2006].
Note that the findings of this thesis are not entirely dependent on the choice of learning algorithm. The random forest algorithm is chosen mainly due to its robust capacity in handling high-dimensional data. Several other common learning algorithms are also used in some of the studies (Paper II and III), including Decision Trees, Support Vector Machines, Logistic Regression, \( k \) Nearest Neighbors, Adaptive Boosting, Bagging, Rule Learner, and Naïve Bayes.

### 4.2.3 PERFORMANCE METRICS

The effectiveness of the learned predictive models can be evaluated by various performance metrics. To complete the hypothesis framework, model A is more effective than model B on criterion C, it is important to choose performance metrics carefully. The ones used in this thesis are accuracy, precision, recall, F-score, area under ROC curve (AUC) and area under precision-recall curve (AUPRC).

Accuracy is one of the most commonly used metrics to evaluate the performance of a predictive model. As the name suggests, it measures how accurate a model is by calculating the percentage of examples that are predicted correctly.

Precision and recall are often used together, as there is a dependence and trade-off between them. The former measures how many examples are correctly predicted as positive among all examples that are predicted as positive; the latter measures how many examples are correctly predicted as positive among all examples that are truly positive. To calculate these, the numbers of true positives (TP), false positives (FP) and false negatives (FN) are needed. A confusion matrix (see Table 4.1) is often used to report them. This is a special contingency table used in machine learning, where columns correspond to the labels of predicted objects and rows correspond to their true labels.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted Yes</th>
<th>Predicted No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Yes</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual No</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>
Precision and recall are then calculated as follows.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}
\]

Depending on the task’s requirements or priorities, it is sometimes more important to optimize the model to obtain either a high precision or a high recall. This can be incorporated by calculating the \( F\)-score, which is the (weighted) harmonic mean between precision and recall, calculated as in Equation 4.4.

\[
F_\beta - \text{score} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \tag{4.4}
\]

The parameter \( \beta \) is used as a weight to prioritize either precision or recall. When such weights are unknown or unspecified, it is common to set \( \beta \) to 1, i.e. \( F_1 \)-score, which corresponds to computing the harmonic mean of precision and recall.

For many tasks involving EHR data, there are fewer positive examples than negative ones, especially when the predictive task is to detect rare outcomes, such as ADEs. When the class distribution of a dataset is not balanced or the error costs among different classes vary, using a metric such as accuracy will lead to results that are biased toward the majority class [Provost et al., 1997]. For learning algorithms that optimize accuracy during training, predicting only the majority class will give a high accuracy if the class distribution is highly skewed, as it often is in many real life problems. For example, among all the patients in a hospital, only a small subset are at high risk of developing a certain disease. In such a situation, AUC and AUPRC are more reliable alternatives.

Area under ROC curve estimates the probability that a model ranks a randomly selected positive example ahead of a negative one. The ROC curve represents a range of trade-offs between the true positive rate and false positive rate, both of which are irrespective of the actual positive/negative balance of the test set. Therefore, AUC is not biased towards the majority class. Random guessing will result in an AUC of 0.5, and a model with an AUC less than or equal to 0.5 is
considered useless. For a binary classification task, the calculation of AUC is straightforward; however, when there are more than two class labels involved, the overall AUC can be estimated by first calculating the AUC for each class versus the remaining classes and then taking the weighted average based on their class frequencies (Equation 4.5).

$$AUC_{\text{multiclass}} = \frac{\sum_{c \in C} \left( \frac{n_c}{\sum_{c \in C} n_c} \times AUC_c \right)}{m}$$  \hspace{1cm} (4.5)

where the AUC for class \(c \ (c \in C)\) with \(n_c\) examples is \(AUC_c\) and \(m\) is the number of classes.

The precision and recall curve has been suggested as an alternative to the ROC curve when the class distribution is skewed [Davis and Goadrich, 2006]. \(AUPRC\) depicts the probability that precision is higher than recall for each recall threshold. Since precision measures how many of the examples predicted as being positive are truly positive, \(AUPRC\) is directly related to how rare the positive class is. It is often preferred when the positive class is rare but more interesting than the negative class [Goadrich et al., 2006].

### 4.2.4 HYPOTHESIS TESTING

Hypothesis testing in this thesis refers to statistical hypothesis testing, where conclusions are drawn using statistical inference [Alpaydin, 2014]. It typically follows a formal process to determine whether to reject a null hypothesis based on sample data according to some level of significance. There are various methods that can be employed for hypothesis testing. Choosing an appropriate one is important since otherwise the reliability and generalizability of the conclusions will be harmed. Two factors are important to determine here: the number of datasets in the sample data and the number of models involved in the null hypothesis.

The sample data in Paper IV consists of a single dataset; in the other papers, multiple datasets are used to test the hypothesis. In Paper IV, McNemar’s test [McNemar, 1947] was applied. This is a non-parametric method that determines whether the marginal frequency of misclassifications between classifiers is equal over a contingency table of mismatched pairs. In the other papers, when two
models are compared to each other, the Wilcoxon signed-rank test [Wilcoxon, 1945] is used to assess the statistical significance. The null hypothesis is that there is no significant difference between the predictive performance of the two models. When more than two models are involved, the Friedman test [Demšar, 2006] is employed for statistical testing of the null hypothesis that the choice of models has no significant impact on predictive performance. A post-hoc test using the Bergmann-Hommel procedure [Garcia and Herrera, 2008] is then conducted to test the pairwise significance, i.e. whether there is a significant difference between each unique pair of models.

4.3 ETHICAL ISSUES

The ethical issues that are relevant to this thesis concern the use of sensitive clinical data and scientific integrity. The sensitivity of the clinical data is protected by law. Such data is typically not available to the public and permission is needed in order to use it. A real EHR database was used in the studies included in the thesis. In this database, patient records are anonymous, i.e., they are deidentified by the hospital by replacing patients’ social security numbers with a random key and by removing all the sensitive information such as names, addresses and phone numbers in the structured fields. It is therefore difficult to make a connection between a health record and the identity of the patient to whom it belongs. This database is stored securely on machines that are disconnected from the Internet in a locked room, which can only be accessed by a limited number of people. To obtain the right to use this database, a confidentiality agreement with the data provider was signed, agreeing on not leaking the data or uploading it online and always using the data in an encrypted format and on machines that are disconnected from the Internet. The published results do not contain any information that can be used to identify patients. The research conducted in this thesis has moreover been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm) with permission number 2012/834-31/5. Given the sensitive nature of the data, it is generally not possible for researchers to publish the data used in each study in order for others to reproduce and validate the reported results. However, for the studies included in this thesis, the contribution is primarily the proposed methods, which can easily be applied by external researchers on data with similar characteristics.
EMPIRICAL INVESTIGATIONS

The proposed methods are evaluated empirically in the context of detecting adverse drug events, which constitute a major public health problem worldwide. The common use case in most of the included studies is to distinguish between patients with an ADE-specific diagnosis and patients with a similar diagnosis that, however, does not indicate that it was drug-induced. A real electronic health record database – the Stockholm EPR Corpus – is used to extract a number of datasets that are used in each study. The various instantiations of the previously described experimental setup for each study are presented, followed by a summary of the main findings from each study with respect to the research questions.
5.1 ADVERSE DRUG EVENT DETECTION

Adverse drug events (ADEs) are injuries caused by a drug or the use of a drug [Nebeker et al., 2004]. The prevalence of ADEs has been increasing over the last decades and constitutes a major public health problem. Worldwide, around 4.9% of hospital admissions are the result of ADEs and this number is as high as 41.3% in some areas [Beijer and De Bley, 2002]. In Sweden, ADEs are the seventh most common cause of death [Wester et al., 2008]. Despite the fact that each drug’s risks are evaluated through clinical trials prior to their release on the market, many unknown side-effects are gradually discovered throughout the later stages of a drug’s life cycle. For example, Vioxx was withdrawn from the market for its doubled risk of causing myocardial infarction [Sibbald, 2004] and the same happened to Cerivastatin, which caused fatal rhabdomyolysis [Furberg and Pitt, 2001]. Therefore, post-marketing drug safety surveillance is necessary and of high importance.

Traditionally, post-marketing drug safety surveillance has mainly relied on analyzing individual case reports. These reports are submitted voluntarily by patients or clinical experts through spontaneous reporting systems (SRSs). Some well-known SRSs are the Yellow Card scheme\(^1\) in the UK, the Adverse Event Reporting System (AERS)\(^2\) by the Food and Drug Administration (FDA) in the USA, and VigiBase by the World Health Organization (WHO)\(^3\), which is maintained at the Uppsala Monitoring Center (UMC) in Sweden. A spontaneous report typically contains the drugs a patient has taken, the adverse events that occurred afterwards and possibly some basic patient information. Although these systems manage to collect a large amount of case reports on ADEs across the world, they have several obvious limitations [Goldman, 1998]. First of all, since they are reported voluntarily, they do not always contain correct information and not all adverse events are fully reported. Secondly, the data in SRSs is mostly based on a single event and not longitudinal, which means that the analysis of such data cannot capture the connections among a series of events. Last but not least, information about patients who take the same drug but do not suffer from adverse events is not available in SRSs, which makes it impossible to estimate the

\(^{1}\)Information available at [https://yellowcard.mhra.gov.uk](https://yellowcard.mhra.gov.uk)

\(^{2}\)Information available at [http://www.fda.gov/Drugs](http://www.fda.gov/Drugs)

\(^{3}\)Information available at [http://www.who-umc.org](http://www.who-umc.org)
incidence rate for adverse events. However, obtaining such estimations are key for the purposes of statistics.

Currently, there are several data sources that can complement SRS data, such as claims databases from insurance companies, EHRs used in hospitals, social media data and biomedical literature [Shibata and Hauben, 2011], among which EHR data has received the most attention. This is not only because EHRs provide a large amount of data across a population over a long time period, but also due to the systematic documentation of healthcare in EHRs that readily allows us to access this longitudinal data. More importantly, the heterogeneous types of data archived in EHRs give us a more complete picture of each patient; the patients’ recorded medical history makes a retrospective analysis possible; also, in contrast to SRSs, information about patients who are not exposed to adverse events is available, which ensures the possibility of estimating incidence rates.

However, like all the other data sources, under-reporting of ADEs is also a problem in EHRs [Hazell and Shakir, 2006]. In EHRs, ADEs are typically reported as part of the diagnoses, which are encoded by a coding system such as ICD-10. These codes are generally assigned by either expert coders or physicians, which means that the quality of assigned codes varies substantially based on the experience and expertise of the coders and physicians. This often results in under-reporting or false reporting of the diagnosis codes [Puentes et al., 2006]. According to the Swedish National Board of Health and Welfare, around 20% of cases contain major errors in the primary diagnosis [Socialstyrelsen, 2006] and about 1.09% of the care episodes in the Swedish patient registry received no primary diagnosis code at all [Socialstyrelsen, 2010]. Therefore, in order to make better use of EHRs for post-marketing drug safety surveillance, it is important to identify cases where ADEs were not properly reported so that they can be taken into account in subsequent analyses.

5.2 STOCKHOLM EPR CORPUS

The data used in all studies of this thesis are extracted from a real EHR system: the Stockholm Electronic Patient Record (EPR) Corpus [Dalianis et al., 2012]. This corpus contains the health records of approximately 1.2 million patients from over 900 healthcare units in the Stockholm region, Sweden. The data
was collected from Karolinska University Hospital in Stockholm from 2009 to 2015. Heterogeneous types of data are available in this corpus, such as basic demographics, diagnoses, drug administrations, clinical measurements, laboratory tests, and clinical notes in free text.

In the Stockholm EPR Corpus, personal information – such as names, addresses, social security number and phone numbers – has been completely removed from the structured data. Each patient is assigned a unique patient ID, which is used to link patient data collected from different departments. Information like diagnoses and drugs is encoded through medical coding systems. Diagnoses are encoded by the Swedish version of ICD-10 and drugs are encoded by ATC. In total, there are 11,623 unique ICD-10 codes, 1,564 unique ATC codes, 1,877 unique clinical measurements and millions of clinical notes. Each patient record contains the patient ID, the time of reporting, clinical events such as an ordered laboratory test, the value of the corresponding event such as the result of the laboratory test, notes of different kinds, as well as various information for administrative purposes.

In this thesis, the extracted patient records contain diagnoses, drugs and clinical measurements, including laboratory tests, as the “raw data” to be used in each study. The proposed methods are then applied to this data to create datasets that can be used as input to machine learning algorithms for building predictive models. For each type of data, patient ID, clinical event name, event reporting time, its value (if applicable) and the value’s unit (if applicable) are included in the raw data. See Table 5.1 for some made-up examples.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Clinical event</th>
<th>Time</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>A11DA01†</td>
<td>2009-04-25 10:33:24</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1234</td>
<td>G444††</td>
<td>2009-04-25 12:17:56</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1234</td>
<td>Blood pressure</td>
<td>2009-04-26 16:42:09</td>
<td>79 mmHg</td>
<td></td>
</tr>
<tr>
<td>5678</td>
<td>I952††</td>
<td>2009-03-14 09:23:11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5678</td>
<td>Body temperature</td>
<td>2009-03-14 09:31:14</td>
<td>36.7 Celsius</td>
<td></td>
</tr>
<tr>
<td>9012</td>
<td>H03BX02†</td>
<td>2009-04-25 14:36:27</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

†: ATC codes encoding drugs. ††: ICD-10 codes encoding diagnoses.

Table 5.1: Made-up examples of raw data extracted from the Stockholm EPR Corpus

4Personal information might however still be present in the clinical notes (free text), which underlines the importance of handling also the anonymized data as highly sensitive.
5.3 STUDY DESIGNS

The empirical investigations were carried out with the use case of distinguishing between patients with an ADE-specific diagnosis and patients with a similar diagnosis that, however, does not indicate that it was drug-induced. To facilitate such ADE detection, a predictive model is learned from the training data with a specific ADE as the positive class label; patients with similar diagnoses make up the negative examples in each dataset.

In EHRs, ADEs are typically reported as part of the diagnosis. Some ICD-10 codes directly indicate that an event is induced by drugs, which can be used for reporting ADEs. There are also codes that suggest a causal relationship between an event and a drug or other substance, as well as codes that do not make an obvious suggestion but are, in practice, often used to report ADEs by physicians. Efforts have been made to summarize the use of ICD-10 codes by physicians for reporting adverse events, where ICD-10 codes are categorized into the following groups that vary in terms of how strong the causal relationship between an event and a drug is [Stausberg and Hasford, 2011].

A.1 A drug-related causation was noted in the ICD-10.
A.2 A drug- or other substance-related causation was noted in the ICD-10.
B.1 The event was denoted as a drug poisoning, thus implying an unphysiological dosage.
B.2 The event was denoted as poisoning by or harmful use of drugs or other substances.
C A drug-related causation was very likely.
D A drug-related causation was likely.
E A drug-related causation was possible.

Among these seven categories, the ones that indicate most clearly a drug-related causation – A.1 (e.g., G44.4 – Drug-induced headache, not elsewhere classified) and A.2 (e.g., I42.7 – Cardiomyopathy due to drugs and other external agents) – were selected for creating training examples in all studies included in this thesis. In the cases of binary classification, each dataset focuses on one specific ADE, where positive examples are patients who have been diagnosed with an ADE-related code from category A.1 and A.2, and negative examples are either patients who are randomly selected from the remaining part of the database (Paper I) or patients
who have been diagnosed with a similar disease to the target ADE (Paper II, III, V, VI, VII, VIII). Similarity is defined by two codes sharing the first three levels, e.g., codes starting with G44 – other headache syndromes, but not G44.4. In the case of multiclass classification, training examples for each class are patients who have been diagnosed with the corresponding ADE-related code, while there is also an additional non-ADE class for which training examples are randomly selected patients without an ADE-specific diagnosis code (Paper IV).

In this thesis, structured EHR data – diagnoses, drugs and clinical measurements – is used for building predictive models. This data is extracted from a defined time window in the patient history. In the studies addressing temporality, the time window is defined to span up to 90 days before the occurrence of the target event; in the other studies, it is defined as the entire patient history available up to the occurrence of the target event. A summary of the datasets used in each paper is presented in Table 5.2, describing the type of data, the number of datasets, classification setting, the number of training examples and the number of features.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Type</th>
<th>Datasets</th>
<th>Class</th>
<th>Examples</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>diagnoses, drugs</td>
<td>30</td>
<td>binary</td>
<td>52-3586</td>
<td>352-3244</td>
</tr>
<tr>
<td>II</td>
<td>measurements</td>
<td>27</td>
<td>binary</td>
<td>41-4671</td>
<td>70-563</td>
</tr>
<tr>
<td>III</td>
<td>all</td>
<td>27</td>
<td>binary</td>
<td>68-3733</td>
<td>1542-8262</td>
</tr>
<tr>
<td>IV</td>
<td>all</td>
<td>1</td>
<td>multiclass</td>
<td>2381</td>
<td>3599</td>
</tr>
<tr>
<td>V</td>
<td>all</td>
<td>14</td>
<td>binary</td>
<td>60-4347</td>
<td>408-3970</td>
</tr>
<tr>
<td>VI</td>
<td>all</td>
<td>14</td>
<td>binary</td>
<td>60-4347</td>
<td>408-3970</td>
</tr>
<tr>
<td>VII</td>
<td>measurements</td>
<td>19</td>
<td>binary</td>
<td>329-9890</td>
<td>1277-6076</td>
</tr>
<tr>
<td>VIII</td>
<td>measurements</td>
<td>19</td>
<td>binary</td>
<td>329-9890</td>
<td>148-513</td>
</tr>
</tbody>
</table>

“all” means that diagnoses, drugs and measurements are used.

5.4 MAIN FINDINGS

Here, the results from the included papers are summarized and presented with respect to how the contribute to answering each related research question, as well as how the included papers are connected to each other. To facilitate the reading, some key terms are first defined as follows:
• **Event type**: the type of clinical activity, such as prescribing a drug, assigning a diagnosis code, or reporting a clinical measurement.

• **Event**: a specific clinical activity, such as prescribing a specific drug, assigning a specific diagnosis code, or taking a specific clinical measurement.

• **Event occurrence**: an event occurring at a specific time point; for instance, a specific drug that is prescribed twice within two days is considered as two different event occurrences.

• **Event value**: the reported result of each event occurrence, for example, an ATC code, an ICD-10 code, or a test result of a clinical measurement.

### 5.4.1 REPRESENTING HETEROGENEOUS TYPES OF DATA

The types of EHR data, or event types, included in this thesis are diagnoses, drugs and clinical measurements\(^5\), each of which has unique characteristics leading to unique obstacles with respect to data representation. Diagnoses and drugs share the characteristic of being encoded with hierarchical classification systems, albeit different ones; clinical measurements differ by typically having multiple occurrences with potentially different values. Paper I targets data representation for diagnoses and drugs, while Paper II handles the representation of clinical measurements. In Paper III, these two types of data are combined to learn a common predictive model\(^6\). Figure 5.1 shows the relationships among the three studies. The problem of representing diagnoses and drugs naturally concerns how to leverage the concept hierarchy of the encoding systems, which is studied in Paper I. The findings of this study are summarized in subsection 5.4.2.

For clinical measurements, which typically have multiple occurrences with different values and often only exist in a small number of health records, the data representation problem focuses on how to address the multiple occurrences

---

\(^5\)The term *clinical measurements* includes both basic measurements, such as body temperature and blood pressure, and laboratory tests.

\(^6\)Learning a common predictive model is in contrast to learning separate models first and then combining them.
Figure 5.1: Relationships of Paper I to III that contribute to representing heterogeneous types of data.

and missing measurements\(^7\). In Paper II, multiple representations of clinical measurements are investigated with respect to addressing the aforementioned two obstacles: \textit{mean}, \textit{standard deviation}, \textit{slope}, \textit{existence} and \textit{count}. The first three representations take into account the actual value from each occurrence to calculate a final score – the averaged value, standard deviation of the values, or how fast the values change over time – to represent a measurement. The last two follow a cruder representation by only indicating whether a measurement is recorded or how many times it is recorded. The results presented in Paper II show that the choice of representation has a significant impact on the predictive performance. Moreover, when many clinical measurements exist only in a small number of health records, using the number of their occurrences is more informative than summarizing their actual values in various ways. It is also shown that combining multiple representations of the same events leads to improved predictive performance, effectively capturing different aspects of the measurements.

\(^7\)Missing refers to the absence of a clinical measurement in a health record, and it merely indicates that there is no value of this measurement for this record in the EHR database. Note that this is different from missing in the sense that the value exists but is not shown or recorded.
To combine these two types of data – clinical codes\(^8\) and clinical measurements – it is not straightforward to find a uniform representation suitable for both due to their different characteristics. In Paper III, a representation is first created for each type of data; the two representations are then combined by fusing the features from each type of data into a common feature space. For detecting certain ADEs, such as \(E273 – \text{drug-induced adrenocortical insufficiency}\), it is shown to be beneficial to combine clinical codes and measurements. However, overall, combining the two only improves the predictive performance compared to using clinical measurements alone, but not compared to using only clinical codes. In this case, the assumption that using more features leads to better performance does not hold. A plausible explanation for this is that using measurements alone leads to rather poor predictive performance and hence makes a limited contribution to the predictive performance when they are combined with clinical codes. As it seems unlikely that clinical measurements are uninformative for detecting ADEs, this suggests that there is room for improvement with respect to representing this particular event type. Representing clinical measurements in this manner fails to account for the latent but valuable temporal information, which motivates subsequent work on incorporating the temporality of clinical measurements (see subsection 5.4.3 below).

5.4.2 LEVERAGING THE CONCEPT HIERARCHY OF CLINICAL CODES

In EHRs, diagnoses and drugs are typically encoded to the most specific level of the corresponding coding system, which provides opportunities to exploit the information that is embedded in the coding systems by merging codes to higher and more general levels. The concept hierarchy of clinical codes is shown to be beneficial for both constructing a feature space (Paper I) and decomposing the predictive task (Paper IV) (see Figure 5.2). The former merges the most specific codes to different hierarchical levels to create a richer feature space, while the latter merges the class labels in form of ADE-related codes to more general levels to decompose the predictive task into a set of sequential steps. Figure 5.3 illustrates the respective use of the concept hierarchy of clinical codes in the two papers.

In Paper I, unique ICD-10 codes, encoding diagnoses, and ATC codes, encoding drugs, are used as features to build predictive models. Besides the original codes that are represented on the most specific level, such as “C10AA01”, the feature

---

\(^8\)Clinical codes refer to diagnoses and drugs, both of which are encoded in EHRs.
Figure 5.2: Relationship of Paper I and Paper IV that contribute on leveraging concept hierarchy of clinical codes.

<table>
<thead>
<tr>
<th>Class Label</th>
<th>A01</th>
<th>A02</th>
<th>A11</th>
<th>A12</th>
<th>A0</th>
<th>A1</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>X01 -&gt; X -&gt; ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X02 -&gt; X -&gt; ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y01 -&gt; Y -&gt; ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y02 -&gt; Y -&gt; ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-ADE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.3: Illustrating the use of the concept hierarchy of clinical codes in Paper I and Paper IV. In Paper I, features in the orange area are created by merging the other features to more general hierarchical levels. In Paper IV, codes that are used as class labels are merged to more general hierarchical levels, so that the predictive task is decomposed into detecting ADE against Non-ADE first, then X against Y, and eventually detecting the specific code.

space is enlarged and enriched with new features, which are created by merging these codes to each higher hierarchical level, such as “C10AA”, “C10A”, “C10” and “C”. The impact on the predictive performance of using features both from each level separately and from different levels in combination is explored. The
EMPIRICAL INVESTIGATIONS 51

The assumption here is that merging clinical codes to different levels captures different information hidden in the data. Overall, using all levels of clinical codes yields significantly improved predictive performance compared to using only the original codes, i.e., the most specific level. This results in a much higher dimensionality; however, it is also observed that the predictive performance can be kept at a high level without employing the more specific levels, for example by using only the top two levels, which at the same time reduces the dimensionality to a large extent since there are many fewer high-level codes (see Figure 5.4).

![Bottom-Up vs Top-Down](image)

**Figure 5.4:** Relative performance of predictive models using features from different levels of the concept hierarchy. The bottom-up strategy adds features on different levels starting from the most specific level to the most general level, while the top-down strategy does the opposite. Higher mean rank means better predictive performance.

In Paper IV, the use of the concept hierarchy is explored from a different perspective, namely the granularity of the predictive task. In this setting, the class label consists of different ADE-related ICD-10 codes and “Non-ADE”, as presented in Figure 5.3. These ADE-related ICD-10 codes can be merged to their shared family level – e.g., “L270” (generalized skin eruption due to drugs) and “L271” (localized skin eruption due to drugs) are merged to “L” (diseases of the
52  CHAPTER 5.

skin and subcutaneous tissue) – and eventually to the coarsest level indicating the existence of any ADE. On the one hand, when the task is to predict a patient having any ADE or not, it is shown to be beneficial to merge the various ADE-related codes prior to building the predictive model. On the other hand, when the task is to detect a specific ADE, instead of a one-step detection, i.e. a single model that takes all the specific ADE-related codes as class labels in a multiclass classification setting, the task can be decomposed into a set of sequential steps: first predicting which patients could have any ADE at all, then predicting the most likely family the ADE belongs to, and finally predicting the exact ADE within this family. It is shown that such a cascading scheme outperforms the one-step detection.

5.4.3 INCORPORATING TEMPORAL INFORMATION

The inherent temporality of clinical events carries valuable information for learning effective predictive models. The temporal information is not only carried by different clinical events, but also by the fact that each event may occur multiple times at different points in time. To incorporate such temporal information in the data representation, it is important to take into account not only the order of the events or occurrences, as well as the time distance between them, but also the event values. Different strategies for incorporating temporal information when learning predictive models from EHR data are explored in Paper V to VIII. The relationships between the papers is illustrated in Figure 5.5.

One option explored in this thesis is to integrate an event’s multiple occurrences into a single-point representation. This means that a single value is created to represent each event. To this end, three integrating strategies are explored in Paper V: bag of events, bag of binned events and bag of weighted events, as illustrated in Figure 5.6. In bag of events, an event is represented as its total number of occurrences during the included patient history; in bag of binned events, an event is represented as its number of occurrences within some pre-specified time windows during the included patient history; in bag of weighted events, an event is represented by the weighted number of occurrences during the included patient history, where the weights are assigned according to the elapsed time from each occurrence of an event to the target event (in this paper, the weight is inversely proportional to the time distance). The first strategy effectively ignores temporality by treating all occurrences equally. The second strategy treats each event’s occurrences from different time windows as different features, where
the number of features can increase tremendously by reducing the size of the time windows. The last strategy treats each single occurrence differently by assigning a temporal weight to each occurrence. It can be noted that the degree of temporality that is taken into account increases successively from the first to the third strategy. In Paper V, it is shown that the last strategy outperforms the first two, especially when involving increasing amounts of patient history in the predictive model. In Paper VI, the temporal weights are pre-assigned in various ways, each of which defines a temporal relationship between a clinical event and the target event. The choice of weighting strategy is shown to have a significant impact on the predictive performance, and which strategy is the best depends, to
some extent, on the dose-dependency of the target ADE\textsuperscript{1}. For example, for a less dose-dependent ADE, each event's earlier occurrences should probably receive less weight compared to the later occurrences, meaning that the occurrences close to the target ADE are more important.

![Diagram of bag of events, bag of binned events, and bag of weighted events]

Figure 5.6: Three strategies of creating single-point representations.

Pre-assigning temporal weights has the limitations of requiring prior knowledge and often failing to capture the exact underlying temporal relationship between an event in the patient history and the target event. To overcome these limitations, in Paper VII, temporal weights are learned in the predictive modeling process. In this case, it is left to the learning algorithm to determine the weight of each occurrence of an event. The learned weights can be applied by (1) using the bag of weighted events representation, and (2) using the weights as probabilities in the feature sampling procedure when building the predictive model. It is shown in Paper VII that, in the former way, learning temporal weights is not more helpful than pre-assigning them. However, this brings up the question of whether it is beneficial to merge different occurrences into a single-point representation. In the latter way of applying the weights, where each event's occurrences are kept separate, it is shown that learning temporal weights significantly improves the predictive performance.

\textsuperscript{1}An ADE being dose-dependent means that it is related to the accumulation of toxins from drugs or medications.
The downside of the single-point representation is that the sequential information is, to some extent, lost, which is likely to be informative with respect to the predictive task. Moreover, the event values are not accounted for in such a representation, where each event is represented as the weighted total number of occurrences, without including their values. Therefore, in Paper VIII, each clinical event's occurrences are treated as a time series, which is then used to represent the event. Since patients often have many clinical events in their health records, this strategy leads to a multivariate time series representation. In addition, the resulting time series can have different lengths within and across features, as well as irregular intervals. For example, patient A has had blood pressure measured 20 times and body temperature measured 8 times, while patient B has had blood pressure measured 15 times and body temperature measured 32 times; both patients have not had these measurements taken on a regular basis but whenever needed. The resulting situation is illustrated in Figure 5.7.

<table>
<thead>
<tr>
<th>ID</th>
<th>C</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>⬤</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.7: A dataset with the multivariate time series representation.

However, such features, i.e. consisting of time series, cannot be used directly by standard learning algorithms for building predictive models. Therefore, in Paper VIII, each time series is first transformed to a symbolic sequence, such as “babcb”, with each symbol representing a value range of the time series, see Figure 5.8.

The resulting symbolic sequences, or strings, can be used directly as features, and are referred to as Original Sequence. This strategy suffers from very diverse strings within each feature, in terms of symbol alignments and number of symbols. This will most likely lead to a poor predictive model due to the difficulty in finding exact matches between the strings, which is problematic since the model makes predictions by finding exact matches between the new examples and the training examples. Two alternatives are hence proposed in Paper VIII: (1) clustering them and using cluster membership as features, which is referred to as Sequence Cluster, and (2) using a distance measure between the original sequences to a random
subsequence as features, which is referred to as Random Subsequence. Prior to model construction, the subsequence with the highest informativeness is selected from a set of randomly generated sequences that are no longer than the original sequences. Both strategies succeed in avoiding diverse strings; however, the quality of the clusters is still influenced negatively to some extent. In Paper VIII, it is shown that the Random Subsequence strategy yields more effective predictive models than the other two and significantly improves the predictive performance compared to the single-point representation that uses time series length as features.
To conclude the thesis, this chapter presents a discussion of the results and conclusions in the included papers with respect to the research questions and previous research, as well as a general discussion on learning predictive models from EHR data and the limitations of the conducted studies. A brief recapitulation of the main contributions of the thesis and conclusions are also provided, followed by suggestions for future research.
6.1 DISCUSSION

This thesis presented and evaluated methods for facilitating the secondary use of EHR data by using machine learning to build predictive models. The research focused on the technical challenges involved in learning predictive models from EHR data and addressed these by investigating issues primarily related to data representation. While EHRs provide large amounts of valuable clinical data, the complexity of the data poses a multitude of challenges for large-scale analysis, including the existence of heterogeneous types of data, the concept hierarchies used for encoding drugs and diagnoses, as well as the inherent temporality of clinical events. In this thesis, empirical investigations of how to learn effective predictive models from EHR data in the presence of these challenges were presented.

How can heterogeneous types of data be represented to allow for effective predictive modeling?

Various representations of each type of data that account for its unique characteristics were investigated. It was shown that combining multiple representations of each clinical event yields improved predictive performance.

In order to build an effective predictive model using heterogeneous types of data, it is fundamental to understand the unique characteristics of each type of data. Diagnoses and drugs share the characteristics of being encoded by hierarchical coding systems, while clinical measurements differ by having actual, mainly numeric, values reported continuously over time. Although a diagnosis or drug can also be reported many times for one patient, there is no value associated to it that may change over time. In this case, if each unique clinical event is to be used as a feature, each diagnosis or drug is linked to a list of the same code; each clinical measurement, on the other hand, is linked to a list of potentially different values.

Based on this information, each clinical code, either encoding a diagnosis or a drug, was represented as a binary feature indicating its existence in a health record, which was also suggested by a pilot study on finding a representative value for each diagnosis and drug [Karlsson et al., 2013]. Here, the focus was mainly on creating features by using the concept hierarchy of coding systems, which is one of the unique characteristics of diagnoses and drugs. The feature space containing codes
on different levels indeed captures different dynamics of the training examples, e.g. patients who share the same disease family but are diagnosed with different specific diseases can be identified. This, in turn, yields improved predictive performance compared to using codes only as they are reported, which are often on the most specific level.

When representing a clinical measurement, the question of how to identify a representative value is an obvious problem. Furthermore, a clinical measurement is typically only reported in a small number of health records, which results in extremely sparse data, with a majority of missing values. By calculating a single score to summarize the group of values, each clinical measurement was represented in different ways depending on the choice of the single score. It is interesting to note that the crude representation merely determining the existence of each measurement in a patient outperformed those taking into account the actual values. This indicates that, instead of treating absent clinical measurements as missing values, explicitly acknowledging their absence is more informative to the predictive model: the fact that a measurement has not been reported is also useful information. Moreover, the feature space is enriched by using more than one score to represent each clinical measurement, effectively capturing multiple perspectives of the data. By doing so, the predictive performance is improved compared to using only a single-score representation.

After the success of building more effective predictive models by combining multiple representations of the same clinical events, the feature space was further enriched by using both clinical codes and clinical measurements. This was achieved by concatenating the different types of data under different representations into one feature set, which is known as feature fusion. There are alternative ways to combine heterogeneous types of data, such as modeling them separately first and then fusing the outcomes from each model, which is known as model fusion. In two of the related papers that are not included in this thesis, we evaluated various strategies for model fusion and compared them to feature fusion [Henriksson et al., 2015a,b]. In the context of detecting ADEs using clinical notes, diagnoses, drugs and clinical measurements from the Stockholm EPR Corpus, it was shown that feature fusion outperformed the investigated model fusion strategies. However, when adding clinical measurements to the feature space containing clinical codes, the predictive performance was not improved overall, as one might have expected. By analyzing feature importance, it turned out that the clinical codes completely dominated the clinical measurements in terms
of their contribution to the predictive model. Therefore, the success of using heterogeneous types of data together was conditional on the performance – and difference in performance – of using each type of data separately. In this case, the reason for the poor contribution of clinical measurements can be two-fold: (1) there are much fewer clinical measurements in the feature space compared to clinical codes; (2) the included representations of the clinical measurements are poor in terms of informativeness. One way to improve the representation of clinical measurements is to take into account temporal information, i.e. the order and time distance among the reported values for each clinical measurement, which motivates the work conducted in the later papers.

**How can the concept hierarchy of clinical codes be leveraged for effective predictive modeling?**

The concept hierarchy of clinical codes was shown to be beneficial for learning effective predictive models from two perspectives: constructing enriched feature spaces and decomposing the predictive task.

In previous research, concept hierarchies have been used to reduce data dimensionality [Singh, 2015] by merging thousands of clinical codes to a more general level, resulting in fewer than a hundred features. The dimensionality is indeed reduced; however, to some extent, information that is only carried by codes on the more specific level is also lost. For example, an ADE is known to be caused by one particular drug, which is only prescribed to very few patients, meaning that using this drug as a feature will be very informative for detecting the target ADE. However, by transforming the ATC code of this drug to a more general level, other patients who are prescribed a similar, but not this particular, drug will also be flagged in the same feature. This is naturally less informative for detecting the target ADE. In this thesis, the concept hierarchy was instead used to create additional features by merging specific clinical codes into more general levels with different granularity. The underlying assumption was that some clinical events are more informative when they are represented by clinical codes on the most specific level, while some provide more useful information when represented by codes on a more general level. Therefore, a feature space containing a combination of clinical codes from different hierarchical levels allows the learning algorithm to choose the most informative level for each clinical event, effectively avoiding the involvement of human experts in the process.
In addition to exploiting the concept hierarchy for enriching the feature space, it was also shown to be useful for decomposing the predictive task. Merging or dividing the ADE-related ICD-10 codes not only improves the predictive performance over using only codes on the most specific level, but also provides flexibility with respect to the predictive task. The latter was achieved by learning an ensemble of models based on different granularity levels of the task. This ensemble model follows a cascading scheme that first detects patients with any ADE, then patients with ADEs from the same family, and finally patients with a specific ADE. In a cascading ensemble, the model in one stage can benefit from the predictive outcomes from the previous stage. In this case, the model detecting the ADE family only needs to consider the patients who are predicted to have any ADE by the previous model. By adopting a cascading scheme, the models in each stage have a more specific task and the training examples – patients in this case – are also less diverse to what they would be in a single, one-step detection model with multiple class labels. Such an ensemble of models would be useful to include in a decision support system, where clinicians could choose to use models from different stages according to their specific needs. In addition, by detecting any ADE or ADEs belonging to the same family, the common characteristics of patients who are diagnosed with these ADEs, if any, can be discovered, which is valuable for alerting clinicians at an early stage that the patient may have suffered from an ADE, thereby improving ADE reporting in EHRs.

**How can temporality be incorporated for effective predictive modeling?**

It was demonstrated how temporality can be incorporated through single-point representations or multivariate time series representations, both of which yielded more effective predictive models than when using representations in which temporality is ignored.

In this thesis, temporality was primarily modeled for multiple occurrences of the same event, not across different types of events. Two ways of incorporating the temporality of clinical events were investigated: temporal weighting and time series modeling. The former assigns weights to each occurrence of a clinical event based on its temporality, while the latter treats those occurrences as a time series. Both of these strategies take into account the sequential information to some extent, which is shown to result in more informative representations than when such information is ignored.
Temporal weights can be obtained by either defining a function that captures a clinical event’s temporal dynamics in relation to the target event prior to the learning procedure, or by letting the learning algorithm identify the temporal dynamics in the learning procedure. Defining a relationship that can reflect the underlying temporal dynamics requires domain knowledge from human experts and is often sensitive to the actual task, e.g., the dose-dependency of the target ADE. This problem can be solved by adopting an alternative way of obtaining weights: learning temporal weights. Furthermore, the manner in which the obtained temporal weights are used sheds lights on how temporality contributes to predictive modeling. In this thesis, two strategies for applying the weights were evaluated: (1) integrating an event’s occurrences subject to their weights; (2) sampling an event’s occurrences using their weights as probabilities. The first strategy treats each clinical event as a bag of its occurrences, which mixes the distribution of each occurrence among training examples; as a result, the impact of each occurrence’s weight is diminished. The high sparsity of the data also makes this strategy problematic. For example, given a clinical event that is reported in only a small number of patients’ health records – resulting in the majority of patients having a zero value for this event – it does not matter how its occurrences are weighted since the small number of non-zeros will most likely be distinguished against all the zeros. However, in the second strategy, the distribution of each occurrence is kept separate; hence highly weighted occurrences have a clear contribution without being contaminated by the lowly weighted ones. The impact of the obtained weights is also more prominent in this strategy as they basically determine if an occurrence will be included in the predictive model at all.

In temporal weighting, the actual value of each occurrence is not used in the predictive model. The dynamics of a clinical event, in terms of values changing over time, is hence not fully captured. By modeling time series, both the order and values of an event’s occurrences are accounted for in the representation. Each clinical event’s occurrences over time are collected to form a time series to represent the event. Each patient has many clinical events reported in the health records and hence is related to multivariate time series. By transforming each time series into a symbolic sequence – a string consisting of a number of symbols – the order of the occurrences is maintained through the alignment of symbols and the value at each occurrence is retained through the choice of the corresponding symbol. The challenge in modeling these time series, in contrast to more traditional multivariate time series problems, is that the length of the time
series varies both within and across features. In some cases the length difference can have a significant impact, meaning that using only the length of a time series to represent each event can result in rather informative features. However, in this case, one is assuming that it is the time series as a whole that may cause the target ADE, which is a fairly risky assumption. Instead, it is more reasonable to assume that it can also be certain sections of the time series that are shared by a group of patients and causing the common ADE. By finding a subsequence that best represents a set of time series within the same feature and using the distance between this subsequence to each time series as features, we hereby confirmed the reasonableness of this assumption.

Overall, it was indeed shown to be important and valuable to incorporate temporality in the data representation. The strategies proposed in this thesis take into account the temporality of clinical events to different degrees, ranging from a weighted bag to multivariate time series. The more detailed temporal information – such as distance on the time span, sequentiality and values changing over time – is incorporated, the more effective the learned predictive models will be.

**General discussion**

In addition to issues directly related to the research questions, there are also some general issues that are pertinent to discuss. One commonly identified challenge for analyzing EHR data is the high dimensionality and sparsity of the data. The reasons for these characteristics of the data are the large amount of clinical events related to patients and the fact that only a small number of patients have been exposed to the same clinical events. These characteristics do, however, not only apply to EHR data; they also pose a common challenge in general machine learning research. There are many studies attempting to address this challenge, such as [Friedman, 1997; Verleysen and François, 2005] in general machine learning and [Karlsson and Zhao, 2014; Henriksson, 2015b; Henriksson et al., 2015b] with respect to EHR data. Although this particular challenge was not addressed directly as a separate research question, some of the proposed representations do in fact lead to reduced dimensionality and sparsity. By merging clinical codes to more general levels in the concept hierarchy, for instance, both dimensionality and sparsity are reduced. Moreover, using unique clinical events as features and incorporating their temporality into the feature representation also greatly reduces the dimensionality and sparsity compared to representing each clinical event by numerous binary features corresponding to the number of its
occurrences. Another positive by-product of the methods developed in this thesis is that they are not conditional on the choice of underlying learning algorithm. This means that any standard machine learning algorithm can be applied with the proposed methods. This hopefully allows the findings of this thesis to reach a broader audience.

Although the developed models are primarily evaluated based on their average performance across multiple datasets with corresponding statistical significance tests, some studies (e.g., Paper III and V) also look at the impact of the models on individual datasets. Even if some models are favored on average, the findings might be different for detecting certain ADEs. There are many factors that can contribute to one method being preferred over another when detecting a specific ADE. One such factor is the dose-dependency of the target ADE. For example, for an ADE with a high dose-dependency, a longer patient history may need to be included in the predictive models; for an ADE with a low dose-dependency, the clinical events that occur close to the target ADE should probably receive much higher weights compared to those that happened a longer time ago. As a result, it is arguable whether a decision support system should adopt a global model for detecting all types of ADEs or a customized model for each specific ADE.

Moreover, certain issues related to how the studies were conducted might be raised. First of all, all the datasets used in this thesis are extracted from the same EHR database: the Stockholm EPR Corpus. This may raise questions on the generalizability of the findings. However, the aim of this thesis is to address the challenges of modeling EHR data, and the challenges identified here are certainly not unique to the Stockholm EPR Corpus. Heterogeneous types of data and longitudinal observations are common characteristics of EHRs in general [Jensen et al., 2012]. Regarding the concept hierarchy of clinical codes, although not every EHR system uses ICD-10 and ATC for encoding diagnoses and drugs, both are promoted by the World Health Organization (WHO) and currently widely adopted worldwide. The proposed methods for leveraging the concept hierarchy of clinical codes are not sensitive to how the hierarchies are created, which means that they should be applicable to other coding systems with concept hierarchies. Moreover, in this thesis, data science research is conducted with a focus on developing and comparing different methods rather than emphasizing absolute predictive performance scores. By applying these methods to other EHR data sources, the absolute predictive performance will most likely change; however,
one may reasonably expect the findings with respect to relative performance to be rather stable.

Another pertinent point is related to the term “effectiveness” that is used in the thesis. Here, it refers to the relative effectiveness compared to some baseline. However, what is to be considered as a baseline is not always obvious when working with EHR data, which means that an external baseline that is well recognized is sometimes lacking. This is unfortunately due to the restricted access to large repositories of clinical data for research purposes, which has resulted in limited research in the exact same domain and fewer methods to which the proposed methods can be compared. In this thesis, a baseline is sometimes a self-proposed, relatively naïve method or the best method observed in a previous study in the same context. This thesis therefore consists of a sequence of studies that continuously aim to outperform their predecessor.

Last but not least, using diagnosis codes that are assigned by physicians during patient visits as class labels can be problematic. We already know that patients are sometimes misdiagnosed in the sense that they are assigned diagnosis codes that do not reflect their true condition. Using these diagnosis codes as ground truth when both generating and evaluating the predictive models is a problem since they have not been verified by clinical experts. In this case, the data used will hence not allow for concluding the true effectiveness of the models, but will only provide an indication of the relative performance of the evaluated methods. The reason for not verifying the diagnosis codes is mainly due to a lack of resources. Here, the studies use multiple datasets with large amounts of training examples; the work required to verify all the diagnosis codes would thus be enormous. Given the limited resources, one option is to extract a small set of data in which the class labels can be verified; however, doing so will result in building models on small data, which is also problematic with respect to reliability. Moreover, in order truly to be able to facilitate the secondary use of EHR data, which is the underlying motivation of this thesis, it is important to exploit the enormous amounts of data available to us. It is therefore not always desirable to create carefully calibrated data for learning predictive models, in particular since this is often not feasible in reality: learning from real, human-generated data involves accepting a certain amount of noise.
6.2 RECAPITULATION AND CONCLUSIONS

Using machine learning to analyze EHRs has great potential to enhance the secondary use of clinical data for improving healthcare. This allows one to build predictive models from historical clinical data in order to make predictions on outcomes of interest. However, given the complexity of EHR data, there are several technical challenges that need to be carefully addressed when building predictive models. The following are three such major challenges that are commonly recognized in EHRs: the existence of heterogeneous types of data, the concept hierarchy used to encode drugs and diagnoses, and the temporality of clinical events. In this thesis, these three challenges are addressed by proposing methods for representing EHR data in a form that can be used by standard machine learning algorithms. Data representation is a key basis for learning effective predictive models: no matter how advanced a learning algorithm is, it relies, to a large degree, on access to high-quality features, in terms of relevance and informativeness, for producing accurate predictions.

When learning predictive models from EHR data, the heterogeneous types of data need to be represented differently given their unique characteristics. Using different data types together does not necessarily lead to more effective predictive models than modeling them separately: it is conditional on the difference in performance of the models using each type of data. To represent diagnoses and drugs, the concept hierarchy of the coding systems is shown to be beneficial for creating a richer feature space, which leads to more effective predictive models compared to using a feature space that only contains clinical codes as they are stored in EHRs, i.e. using only the most specific level in the concept hierarchy. The concept hierarchy is also useful for decomposing the predictive task into a number of sequential steps by using codes on different levels as class labels. When representing clinical measurements, the extent to which the temporal information among their continuous occurrences can be taken into account plays an important role. This thesis proposes two strategies for incorporating the temporality of clinical events into data representation: (1) creating a single-point representation where temporality is used to weight each event’s occurrences; (2) representing each event as a time series containing its occurrences over time. Both strategies yield more effective predictive models compared to representations that ignore the temporality of clinical events.
Overall, the main contribution of the thesis is to propose and evaluate data representation methods that improve the effectiveness of predictive models learned from EHR data. In particular, methods for representing heterogeneous and longitudinal EHR data are developed. These methods succeed in taking into account the unique characteristics of different types of data, including concept hierarchies and temporality, in their representation, which allows useful information embedded in the data to be captured by the learning algorithms. Another advantage of the proposed methods is that they can be used in combination with any standard machine learning algorithm and are not restricted to a specific task, even if they were here evaluated in the context of detecting ADEs. This, to a large extent, facilitates meaningful secondary use of EHRs. Effective data representation methods are key for enhancing large-scale analysis of EHRs, which, ultimately, will become a cornerstone of data-driven initiatives for improving healthcare.

6.3 FUTURE DIRECTIONS

There are several potential future directions that can be followed based on the work presented in this thesis. The most obvious one is perhaps to take a step further on addressing the challenge posed by heterogeneous types of EHR data by exploring different strategies for exploiting these in combination when learning predictive models. Although modeling the heterogeneous types of data separately has been studied to some extent, it is natural to make an effort to combine them such that a richer perspective of patient trajectories can be captured. Combining structured EHR data and the narrative notes has been touched upon [Henriksson et al., 2015a,b]; however, only rather straightforward fusion strategies were explored. Using different types of data can be achieved in other ways than fusing features or models extracted from each type of data. The key is to find a way to account for the hidden, more complex, relationships between different types of data. On the one hand, if a single feature space is to be created from different types of EHR data, the fact that these complement each other in giving a more complete picture of a patient – not only in terms of reported content but also time – needs to be taken into account. For example, the explanation for assigning a diagnosis code may be described in the clinical notes, where the use of a drug a week ago is mentioned; the current observed ADE is caused by the interaction between this drug and another drug that is prescribed at the same time, but not by the drug that
is prescribed recently to treat the symptom of the assigned diagnosis. Here, the order of events from the notes, diagnoses and drugs needs to be accounted for in the predictive model in order to detect the ADE accurately. On the other hand, if one model is to be learned from each type of data, a final predicted outcome needs to be extracted based on the outcomes of these models. In addition to merging their predicted outcomes by majority voting or manipulating the probabilities, which is commonly done, there are alternatives for producing the final predicted outcomes by combining the models in various other ways [Yang et al., 2003]. There is hence space for exploring more advanced methods for combining different types of EHR data.

Another potential direction would be to link EHR data to other data sources, such as data from molecular-level and genome data. Including other data sources is a natural option in order to further improve the richness of data representation and to allow for more effective predictive modeling. Besides using the concept hierarchy of ATC codes to better represent drugs, for instance, the fingerprint of each drug on a molecular level can also be included to enrich the feature space. In addition, even though the phenotypical profiles provided by EHRs already contain a vast amount of information for each patient, linking them to biobanks and genetic data will not only extend the feature space for predictive modeling, but also deepen our understanding of patient variations for creating better case and control groups.

Finally, it would be interesting to explore deep learning techniques to learn the data representation automatically. Deep learning has recently become extremely popular for its robust capacity in modeling complex data such as natural language, images and speech [Linggard et al., 2012]. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. The learned data representation is typically very abstract and hidden in one of the middle layers, which allows no external control over the representation. The learning procedure is fully automatic, taking raw data as input and producing predictions as output. Therefore, it is not designed merely for data representation; however, if interpretability of the learned features is not required, deep learning may be a promising solution.
REFERENCES


Ehtesham Iqbal, Robbie Mallah, Richard George Jackson, Michael Ball, Zina M Ibrahim, Matthew Broadbent, Olubanke Dzahini, Robert Stewart, Caroline


