Predicting clinical outcomes via machine learning on electronic health records

Awais Ashfaq

LICENTIATE THESIS | Halmstad University Dissertations no. 58

Predicting clinical outcomes via machine learning on electronic health records

Halmstad University Dissertations no. 58
ISBN 978-91-88749-24-6 (printed)
ISBN 978-91-88749-25-3 (pdf)
Publisher: Halmstad University Press, 2019 | www.hh.se/hup
Printer: Media-Tryck, Lund

Abstract

The rising complexity in healthcare, exacerbated by an ageing population, results in ineffective decision-making leading to detrimental effects on care quality and escalates care costs. Consequently, there is a need for smart decision support systems that can empower clinician's to make better informed care decisions. Decisions, which are not only based on general clinical knowledge and personal experience, but also rest on personalised and precise insights about future patient outcomes. A promising approach is to leverage the ongoing digitization of healthcare that generates unprecedented amounts of clinical data stored in Electronic Health Records (EHRs) and couple it with modern Machine Learning (ML) toolset for clinical decision support, and simultaneously, expand the evidence base of medicine. As promising as it sounds, assimilating complete clinical data that provides a rich perspective of the patient's health state comes with a multitude of data-science challenges that impede efficient learning of ML models. This thesis primarily focuses on learning comprehensive patient representations from EHRs. The key challenges of heterogeneity and temporality in EHR data are addressed using human-derived features appended to contextual embeddings of clinical concepts and Long-Short-Term-Memory networks, respectively. The developed models are empirically evaluated in the context of predicting adverse clinical outcomes such as mortality or hospital readmissions. We also present evidence that, surprisingly, different ML models primarily designed for non-EHR analysis (like language processing and time-series prediction) can be combined and adapted into a single framework to efficiently represent EHR data and predict patient outcomes.

To the ONE who gives me life...

Acknowledgements

I would like to express my deepest gratitude to my principal supervisor Slawomir Nowaczyk for welcoming me into his research team and providing me with an opportunity to explore science, participate in national and international research venues and present ideas orally and in writing. His mentorship and constructive criticisms have significantly influenced, for good, the way I understand and approach research work.

I also like to sincerely thank Markus Lingman from the Halland Hospital for architecting the bridge between medicine and machine learning for me. His firm belief in improving care-delivery through cross-disciplinary agile teams and data-driven solutions has always been a great motivation for my work. I am also grateful to both Slawomir and Markus for facilitating my collaboration with several external researchers that has flavoured my research experience with valuable cultural and scientific diversity.

I like to acknowledge the efforts of Anita Sant'Anna and Jens Lundstrom who co-supervised my work and helped formulate my research scope. I also like to acknowledge all the support (either relating to research or general matters) that I received from my colleagues in the lab: Pablo, Yuantao, Shiraz, Hasan, Jennifer, Ece, Maytheewat, Suleyman, Rebeen, Kevin, Alex, Sepideh, Rafiq, Taha and Naveed. I am grateful for the occasional fun we had in the city or at each other's places and I hope it continues in the future. I also appreciate the timely and effective research support that I have enjoyed from my seniors in the lab and hospital: Antanas V., Magnus C., Roland T., Miltiadis T., Thomas W., Stefan L. and Stefan B.

Last but not least, I can't thank my parents, wife and brothers enough for their never-ceasing moral support and appreciation for what I do. With absolute certainty, all that is good in my life rests on the foundation of their hard work, sacrifice, and love.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **Data profile: Regional Healthcare Information Platform**
Awais Ashfaq, Stefan Lönn, Håkan Nilsson, Jonny A. Eriksson, Japneet Kwatra, Zayed M Yasin, Jonathan E Slutzman, Thomas Wallenfeldt, Ziad Obermeyer, Philip D Anderson, Markus Lingman. **International Journal of Epidemiology**, (2019) *submitted*.

PAPER II: **Training machine learning models to predict 30-day mortality in patients discharged from the Emergency Department**
Mathias C. Blom, Awais Ashfaq, Anita Sant'Anna, Philip D.Anderson, Markus Lingman. **BMJ Open**, (2019) *submitted*.

PAPER III: **Readmission prediction using deep learning on electronic health records**
Awais Ashfaq, Anita Sant'Anna, Markus Lingman, Slawomir Nowaczyk. **Journal of Biomedical Informatics**, (2019) *submitted*.

Paper contributions

Paper I	I contributed to the study design, performed the analysis and wrote the majority of the manuscript.
Paper II	I contributed to the study design, performed the experiments and wrote parts of the manuscript.
Paper III	I came up with the idea, contributed to the study design, performed the experiments and wrote the manuscript.

Contrary to all other text in this thesis, this section is written in the 'first person' format to highlight personal statements about the authorship of the papers

Contents

List of Figures

List of Tables

1. INTRODUCTION

Predictive modelling using comprehensive Electronic Health Records (EHRs) is envisioned to improve quality of care, curb unnecessary expenditures and simultaneously expand clinical knowledge. This chapter motivates the big picture on why the application of machine learning on EHRs should no longer be ignored in today's complex healthcare system. As promising as it sounds, clinical data comes with a multitude of data-science challenges that impede efficient learning of predictive models. The challenges relate to data representation, temporal modelling and data bias impact which, in turn, drive the research presented in this thesis. In particular, one contribution of this work is a cost-sensitive Long-Short-Term-Memory (LSTM) network for outcome prediction using expert features and contextual embedding of clinical concepts.

AI is about human empowerment

Doctor: I might lose my job because of you.
AI: I am not the threat. Doctors who use me are.

1.1 Motivation

The arc of history drawn by Sir Cyril Chantler in 1999 is increasingly clear; *"Medicine used to be simple, ineffective, and relatively safe. Now it is complex, effective, and potentially dangerous"*.

The complexity of modern medicine is primarily driven by the complex human biology which is subject to nearly constant change within physiological pathways due to a series of gene/environment interactions. As a result, medical knowledge is expanding rapidly [1]. For instance, International Classification of Diseases (ICD-10) specified over 68,000 diagnoses (five times the size of ICD-9) and the list keeps growing as we await ICD-11 [2]. In order to cure or alleviate patient sufferings, clinicians practice thousands of drugs and therapies. Simultaneously, the demand of healthcare is rising with escalating and ageing population [3]. Moreover, studying the variabilities in genes, environment and lifestyles of humans have progressed the idea that medical care should be customised to individual characteristics and care patterns [4]. Thus, adding another wave of information including comprehensive patient-specific factors which may easily number in thousands. While the present digital era has equipped modern medicine with effective tools to store and share information, the ability to assimilate and effectively apply the unprecedented amount of knowledge generated in medicine far exceeds the capacity of an un-aided human mind [5].

In addition to the cognitive overload that impedes effective application of medical knowledge, there also exist situations where relevant information for clinical-decision making simply does not exist. A widely accepted source of clinical knowledge and evidence comes from randomized control trials (RCTs) [6]. These are quantitative and comparative experiments to investigate the effect of one or more interventions in a random population while minimizing the effect of confounding factors. However, the guidelines learned from conducting well-designed RCTs are often not reflective of the 'average' patient. This is because RCTs follow strict exclusion criteria due to comorbidity, polypharmacy, pregnancy, history of noncompliance and more [7]. Since such patients are among the frequent consumers of healthcare resources; the findings of RCTs cannot be easily generalised to commonly seen patients in care centres. Put differently, it means that clinicians often need to make decisions about patients with limited guidance from medical knowledge.

Therefore, it is not surprising that many clinical decisions are not optimal in terms of patient care and costs. Patients are often prematurely discharged to

homes from hospitals and are soon readmitted [8]. Emergency care patients are often unnecessarily admitted to hospitals [9]. Patients are often subjected to painful and costly surgeries, yet some die soon after the procedure [10]. Similarly, misdiagnoses, over-diagnoses and unnecessary medical tests are increasingly common [11]. All these pose a negative impact on patient's wellbeing, care quality and escalate costs. Since clinical decisions deal with human lives, we want to be as certain as possible about outcome before making a decision.

Precise and timely insights into (or predicting) individual patient outcomes can facilitate clinical decision support in favour of better patient outcomes. For instance, patients at high-risk of 30-day readmission may benefit from personalized discharge planning, care counselling or be considered for home nurse visits before leaving the hospital. Machine Learning (ML) can provide future insights with a significant degree of precision when applied on clinical data. ML is a scientific discipline that focuses on algorithms that learn complex patterns in historical data. In a clinical setting it can be used, among others, to predict future outcomes of interest [12]. Often times, we are interested in predicting adverse outcomes (AO) so that necessary actions are taken to avoid them (if avoidable) or prepare for them. AOs include, but are not limited to, disease onset, hospital readmission and mortality.

The application of ML in healthcare is widely anticipated as a key step towards improving care quality [13]. A boon to this anticipation is the widespread adoption of Electronic Health Records (EHRs) in the health system. In Sweden, EHRs were introduced in the 1990s and by 2010, 97% of hospitals, and 100% of primary care doctors used them for their practice [14; 15]. EHRs are real-time digital patient-centred records that allow secure access to authorized care-providers across multiple healthcare centres when required. The structure of EHR consists of temporally ordered patient visits that carry (i) clinical information, such as, patient symptoms, diagnoses, treatments, lab results, medications; and, (ii) relevant demographics about the patient (age, gender etc.), his/her visit (date, type etc.) and the care-provider (age, qualification etc.). Additionally modern EHRs are built to bridge disparate data sources such as health registries, billing, claims, patient-generated reports, genomics and more to facilitate an in-depth understanding of the healthcare system in general and the patient's care process in specific [16; 17]. This ongoing digitization of healthcare generates unprecedented amounts of clinical data, which when coupled with modern ML tools provides an opportunity to expand the evidence base of medicine and facilitate clinical decision process.

One key step in building ML models is to extract a set of features or predictor variables from the input data. Predictors can be any information in the input data like age, gender, haemoglobin level, diagnose and more. A common approach is to have domain expert select features based on relevance to the prediction task. This step is often referred as Feature Engineering in ML. While domain based feature extraction is widely used in building prediction models, it scales poorly because prediction performance largely hinges on input features which are often task specific. For instance, "blood pressure" is a more relevant feature for predicting "heart failure" rather than predicting "bone fractures". Feature engineering is also oftentimes referred as a 'black art', demanding creativity, insights or luck. Moreover, the number of potential predictors in EHRs may easily number in thousands, yet traditional approaches rely on considering limited number predictors. A recent review investigated 107 articles on EHR driven prediction models and the median number of input variables used were found to be only 27 [18]. Put differently, a vast majority of information is not included in the prediction models, thus, discarding the opportunity to extract new patterns of relevance and generate knowledge.

Unsupervised feature learning and deep learning have outperformed the limitations of feature engineering in many data analytic applications like speech and language processing and computer vision [19]. Deep learning contributes by automatically learning features that are best suited for the prediction task at hand. Put differently, it shifted the paradigm from expert-driven features to data-driven features for prediction models. Given the challenges of feature engineering from EHRs, it is not surprising that deep learning approaches are highly successful for EHR driven prediction modelling [13]. However, EHRs offer a unique set of challenges for deep learning research due to their complex structure. The next section elaborates some common data-science challenges inherent to EHRs which, in turn, drive the research presented in this thesis.

1.2 Problem statement

EHR data does not have a clear spatial structure (like pixels in an image) or sequential order (like natural language or time series data). Rather it constitutes a heterogeneous mix of different data types specifying different clinical concepts and demographics about the patient and the care system. Summarizing and representation these clinical concepts and patient data in EHRs is a cornerstone of building prediction models via ML. Data in EHRs can be broadly grouped into structured and unstructured data. In this thesis we focus on structured EHR data. Structured data means documented patient information using

a controlled vocabulary rather than free text, drawing or sound. For instance, *'hypertension'* can be recorded as *'high blood pressure'* or *'>140/90 mmHg'* etc. but is instead recorded as *'I109'* according to a standardized schema called the International Classification of Diseases (ICD-10). Similarly there exist other schema like the Logical Observation Identifiers Names and Codes (LOINC) to code laboratory exams, Anatomical Classification System (ATC) to code medications and more.

Hitherto, we specify some key problems in modelling structured EHR data for ML research.

- *Heterogeneity:* EHRs store various types of data including patient's and care provider's demographics, patient's diagnoses, procedures, symptoms, lab exams and results, prescribed medications and more. The data is heterogeneous both in terms of clinical concept and data type. For instance, multi-range numerical quantities such as lab results; date time objects such as visit dates and time; categorical data such as diagnostic codes or visit locations and more. This clearly distinguishes EHRs from other data sources that have homogenous raw inputs like fixed-range pixel values in images or alphabets in natural language. This mixed-type EHR data drives an interesting research field of how best to combine them for learning prediction models.

- *Dimensionality:* Another inherent challenge with EHR data is the dimensionality of clinical concepts. ICD-10 specifies over 68,000 unique diagnoses. Similarly there exist thousands of different procedures, labs and medications in medicine. The problem of high-dimensionality are often coined as *the curse of dimensionality* [20]. The terms captures that increasing dimensionality comes along with, (i) an increase in model complexity by adding more parameters, (ii) the need for more training samples to avoid model overfitting since the feature space gets sparser, and (iii) the inefficiency of common data organization strategies since the samples get more dissimilar. Of note, there is no clear definition in the literature for 'high dimensionality'. In some situations, data sets with as few as ten features have also been referred to as being high dimensional [21].

- *Temporality:* EHRs include time-stamped sequences of measurements (clinical visits) over time which contain important information about the progression of disease and patient trajectory over the care period. The sequences are irregularly sampled. Both the order of clinical events and the time difference between events are valuable pieces of information for

learning prediction models. However, how best to encode the temporal information into EHR driven predictive modelling is unclear, primarily because this a relatively new challenge that is not common in many other domains.

- *Data bias:* EHR driven prediction models often suffer from Class Imbalance Problems (CIP). A dataset is referred to as *skewed* or *imbalanced* if some classes are highly under represented compared to others. General prediction algorithms assume that training sets have evenly distributed classes which – in case of skewed datasets – biases the algorithm towards the majority class. As a result, the distribution of the minority class is not adequately learned [22]. CIPs are troublesome in fields where correctly predicting the minority class is often more significant than the other. For instance, in the clinical domain, a false-negative HIV result prior to renal dialysis can be catastrophic [23]. A preferred prediction model is often the one with a high precision on the minority class with a reasonable precision on the majority class. Put differently, it means relaxing the decision boundary of the classifier in favour of higher sensitivity or recall at the cost of a low precision. Additionally, some recent studies have shared concern towards understanding the complexities and limitations of EHR data when learning predictive models [24–26]. The primary purpose of EHRs is to effectively store and share patient information across different care providers, and not clinical analysis or research. EHRs are, thus, populated by Routinely Collected Data (RCD) which introduces new challenges like the variability in completeness of data among patients over time resulting in missing information. Data completeness often depends on the underlying condition of patients, care-provider and workplace as well as varying care policies in the healthcare system. These factors might or might not affect the performances of different prediction models. Either way, exploring the possible sources of biases in EHRs and investigating and quantifying the effect of those on prediction model performance remains unclear.

Of note, heterogeneity is, to a large extent, entwined with the dimensionality challenge. Thus they are addressed together in the research questions explained below.

1.3 Research questions

The research questions in this thesis are presented from a clinical and data science perspective. From a clinical perspective, the thesis aims to answer:

What is the probability of a given outcome for a patient at a specific time given data available at that time?

In this thesis we define outcome risk as mortality or readmission risk of patients at a given time. Death is a bad outcome in healthcare with an exception for someone suffering from a terminal illness where death is the only option left. Unplanned readmissions are bad because they disturb the normality of the patient lives and contribute significantly to healthcare costs. Both mortality and unplanned readmissions are also considered as a proxy by authorities (like the Centre for Medicare and Medicaid Services in the US) to measure care quality and hospital reimbursements [27]. This is because these outcomes might be related to improper treatment, clinical errors, premature discharge or more [28].

Correspondingly, from a data science perspective, the thesis aims to answer:

- ***How can heterogeneous clinical data be represented for EHR driven prediction models?***

  A patient record in an EHR is described by a sequence of visits and each visit includes demographics and clinical information of the patient. Demographics include patient age, gender, visit type etc. The clinical information includes concepts like diagnoses, procedures, medications, labs and more. Here, the goal is to combine and represent the information about each visit in a machine-friendly way to enable learning of predictive models. Put differently, we learn numerical representations of patient visits.

- ***How can temporal information be incorporated in EHR driven prediction models?***

  Temporal dynamics in EHRs refer to the sequence of visits in this thesis. In other words, we move from visit to patient representation now. Of note, we do not consider the actual time gaps between visits at this stage. The goal, here, is to capture the order of patient visits and events in the EHR and include this information when building prediction models.

- ***What are the possible sources of bias in EHR data? How can we quantify the effect of those biases on the performance EHR driven prediction models?***

We primarily focus on class imbalance in this thesis since EHRs are often highly skewed records with very few samples containing the target outcome of interest. The goal is to tailor the model training process to learn the distribution of the minority class as well. In the later stage of the thesis, we expand on exploring and quantifying the effect of other sources of biases in EHR data.

We detail the contribution for each research question in section 1.5.

1.4 Research approach

This thesis embraces the positivist research approach which is based on the empiricist view of epistemology - obtaining knowledge through observation and experimentation [29]. This approach synchronizes with the deductive nature of research in machine learning applications where a hypothesis is developed and tested through experiments. Here, the hypotheses are often drafted in a comparative form: method A is better than method B in terms of an evaluation metric. The methods include, but are not limited to, data representation schemes, model choice and optimization techniques. The observations in the experiment are predictions often on a hold-out dataset or test set. The predictions are then compared to the ground truth to calculate an evaluation metric like accuracy, precision, recall, ROC-AUC etc. Finally the evaluation metric is used for statistical inference and hypothesis testing.

Despite the research being primarily deductive, it is also - to some extent - exploratory and descriptive. As discussed earlier, EHR analysis comes with a multitude of interesting challenges, and the exploratory phase of this research was intended to establish priorities given interest, utility and available resources. Based on the findings in the exploration step, descriptive studies were conducted to understand and quantify the problem at hand in particular from a clinical perspective. These include understanding the complexity of the EHR data and care patterns of Congestive Heart Failure (CHF) population. For instance, unscheduled readmissions are a hallmark of CHF population with readmission rates approaching 30% within 30 days from hospital discharge resulting in an estimated cost burden of 70 million SEK per annum. This in turn provides a strong clinical motivation to invest on new methodologies to predict early of risk of readmission and trigger appropriate readmission-preventing interventions if necessary. Later, while attempting to build prediction models using EHR data, emerge other challenges - often not thought of in the beginning - like evaluating the impact of data biases on the model performance.

Fig 1.1 provides an overview of the research conducted so far. Hexagon represents the exploratory stage of the research. Papers I and pre-print IV cover descriptive studies mainly from a clinical perspective. Rounded rectangles mention data science challenges together with attempted solutions in rectangles. Papers II and III are completed deductive studies using ML techniques. Pre-print V is work in progress.

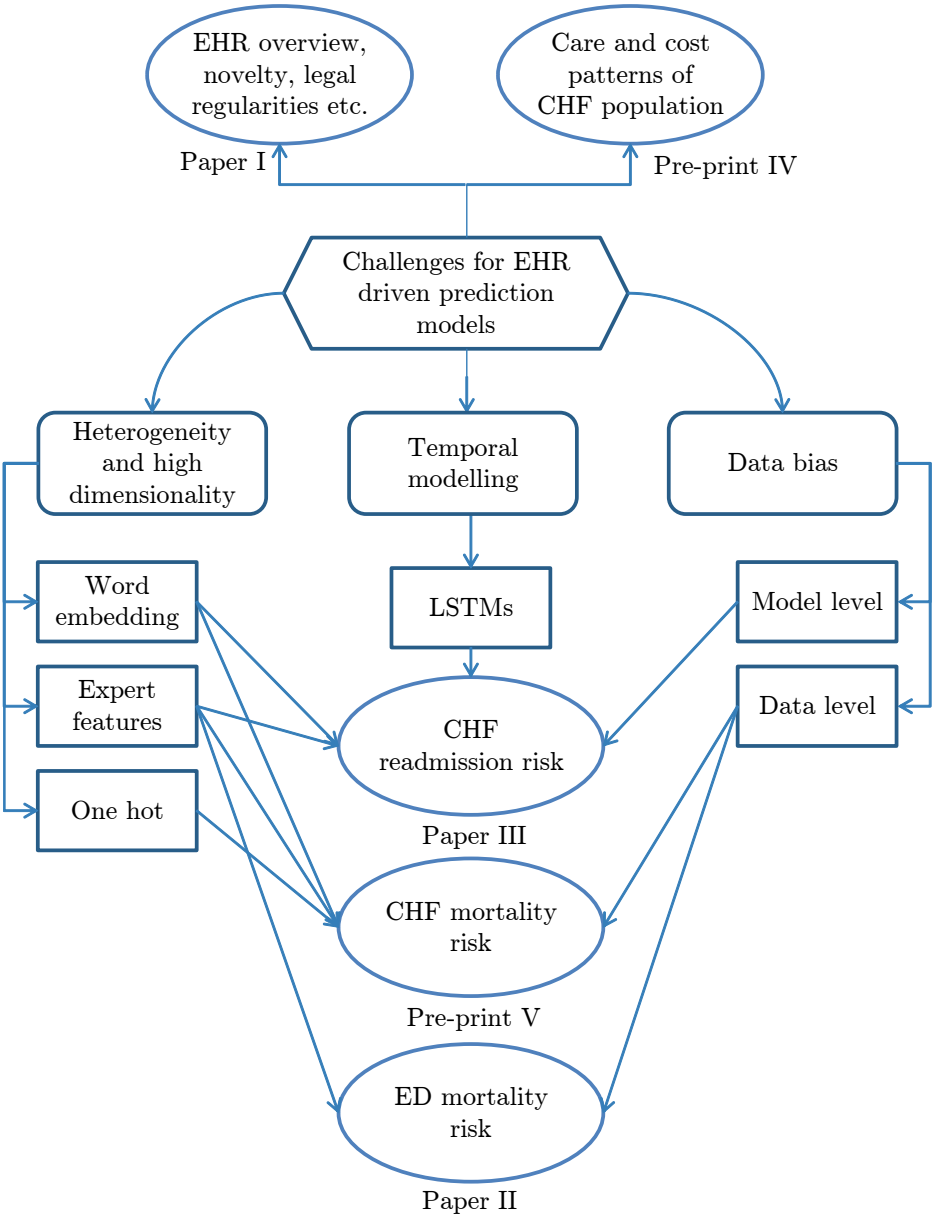

Figure 1.1: Graphical summary of research conducted so far. Each conducted research will be described in the corresponding papers.

1.5 Contributions

Thesis contributions are explained from a clinical, economical and data science perspectives. From a clinical perspective, decision support systems for

predicting mortality (Paper II) and 30-day readmission (Paper III) are developed with state-of-the-art prediction performance. This allows patients at risk of an adverse outcome to benefit from personalized discharge planning, care counselling or be considered for home nurse visits - in order to avoid or prepare for the adverse outcome.

From an economic perspective, we show (Paper III) that a high precision model will have significant financial savings if appropriate interventions are targeted to high-risk patients. This not only facilitates precision resource utilization in hospitals, but also encourages policy-makers to focus on real-time implementation of such decision support systems in clinical practice.

Contribution summary in light of the data science research questions is presented below.

- ***Representation of heterogeneous clinical concepts:***

  The presence of standard coding schema of clinical concepts plays a pivotal role in facilitating the representation of clinical concepts in EHRs. In general the codes are very granular with varying specificity levels. However, the schemas follow a standard hierarchy that can be exploited to group similar concepts. This, to some extent, solves the high dimensionality problem, reduces information overload and provides a common specificity level across the data.

  The lists of clinical codes present in a visit profile of a patient in EHR exhibit strong similarity to common Natural Language Processing (NLP) applications. Here, codes can be treated as words and visits as sentences. Similarly, a patient can be treated as a document or a sequence of sentences (visits). Following the text analogy, NLP techniques like word2vec can be leveraged to not only reduce the dimensionality of the code space but also preserve latent similarities between clinical concepts.

  A key limitation of NLP inspired embedding techniques is that they do not account for rare words in the vocabulary. Rare clinical codes like cancer or HIV can have significant importance if present in the patient profile. Thus, one promising solution is to complement the embeddings with domain expert features and we show (Paper III) that it helps improve the model performance.

While deep learning has empowered automatic feature selection from huge data sets in recent years, the utility of good old expert driven feature selection should not be ignored - particularly in clinical applications. Deep learning algorithms demand huge amounts of data which is, often times, challenging to access in healthcare due to stringent data protection laws and ethical concerns [30]. Thus for simple prediction problems, using supervised definition of feature space from published literature or via directed acyclic graphs as agreed upon by a committee of experts can provide state-of-the-art performance (Paper II).

- ***Incorporating temporal information in EHR driven prediction models:***

  We show that an important piece of information in EHR is embedded in the sequential trajectory of patients (Paper III) which is consistent with previous studies [31; 32]. A patient can be represented as a variable length sequence of fixed size vectors where the vectors correspond to visit information. For such data structures, Long Short Term Memory (LSTM) networks are promising tools due to their inherent nature of handling variable length input sequences. Since visits in EHRs are irregularly sampled, further investigations are in progress to append the time-gap information at each visit. From a clinical perspective, recent past events are often more influential in determining patient outcomes than much older events and, in general, LSTM's do not account for actual time-gaps between sequences.

- ***Addressing the class imbalance problem:***

  We show that adjusting for class imbalance strikes a reasonable balance between sensitivity and specificity values and does not contribute in terms of ROC-AUC (Paper III). This is inconsistent with [33] that proposed a cost-sensitive deep learning model for readmission prediction. We suspect that the improved model performance in that study was because of using both static and continuous EHR features for prediction and not because of cost adjustments. However, it requires further investigation.

1.6 Disposition

The remainder of the thesis is organised as below. Descriptive studies (Paper I and pre-print IV) are summarised in chapter 2. Chapter 3 targets predictive modelling and includes details on developed models and results on clinical applications (Paper II and III). Finally, chapter 4 highlights ongoing work (pre-print V) and future directions, followed by some concluding remarks.

2. BACKGROUND

Accurate and comprehensive datasets are necessary for machine learning research. This chapter presents the Regional Healthcare Information Platform developed in Sweden which was primarily used as a data resource in this thesis. A novelty of this platform is that it encapsulates a 360 degree data view of a patient covering clinical, operational and financial information from all public care facilities in Halland, Sweden. Put differently, an advantage over traditional EHRs is that care data off the hospital radar is also archived, such as visits to primary care, outpatient specialty, emergency care and even pharmacy pick-ups.

Demonstrating the comprehensiveness of the data platform, we conducted a detailed observational study of patients with CHF in Halland, Sweden. It was found that the CHF population (2.0% of the total region population) are heavy consumers of healthcare resources, in particular inpatient care costing €30M or 15.8% of the region's total inpatient care expenditure. Inpatient care is the care of patients whose condition requires admission to a hospital. Data from this specific population was later used in machine learning studies to predict hospital readmission risk. This will facilitate timely initiation of relevant interventions to reduce unnecessary admissions and curb care costs.

2.1 Data profile

In the previous chapter, we motivated the utility of ML on EHRs to facilitate decision support and reduce cognitive overload on physicians. Given that, the importance of a comprehensive EHR in itself should not be ignored. Advancements in medical knowledge and guidelines have progressed a paradigm shift in healthcare: from care in a single unit to care across multiple units with varying but specialized expertise. It is often referred as *fragmentation* in medicine. Thus, patient care (and of course data) is split along multitude of different facilities and computer systems and integration of this information into a single system faces numerous challenges, primarily from an organizational perspective [34; 35]. These include privacy and security concerns, lack of acceptable standardized data formats, use of proprietary technologies by disparate vendors, costly interface fees and more. As a result, care-providers are impeded from accessing complete datasets and thus unable to understand all aspects of the patient health journey. It also results in redundant clinical work and procedures adding unnecessary cost on the care system [36].

Paper I describes the Regional Healthcare Information Platform in Halland, Sweden. It is a novel healthcare analysis and research platform covering pseudo-anonymized clinical, operational and financial data on over 500,000 patients treated since 2009 in different public care facilities in Halland, Sweden. Tab 2.1 shows a brief overview of the data content.

Population (N), % of total population	Alive 514 986 (94.4%)		Deceased 30 666 (5.6%)	
Gender distribution	Male 49.2%		Female 50.8%	
Visit count (N), % of total visits	Primary care 9 482 484 (59.0%)	Specialty outpatient care 5 457 707 (33.9%)	Inpatient care 400 339 (2.5%)	Emergency care 742 007 (4.6%)
Care centers (N)	Primary care 24	Hospitals 3		Emergency care 2

Table 2.1: Population statistics in the regional information healthcare platform in Halland, Sweden. January 2009 - October 2018

Similar to most EHRs, the frequency of data collection in the platform hinges on individual patient needs and is influenced by age and underlying morbidity. The data, thus, reflect real-world practice. Moreover, data are time-variant (updated monthly).

Diagnoses and procedures in the regional platform are encoded in accordance with the Swedish version of the 10th edition of International Classifica-

tion of Diseases (ICD). Medications are encoded using the Anatomical Therapeutic Chemical (ATC) classification system. Both ICD-10 and ATC are hierarchical alpha-numeric encoding schemes where the specificity of the disease, procedure or medication increases as we move from left to right. Lab exams in the regional platform are encoded in a proprietary format.

The platform is designed to operate within Swedish and EU laws (General Data Protection Regulation 2916/679) respecting confidentiality and privacy pertaining to patient and corporate data. We detail further description of the platform in Paper I.

2.2 Congestive heart failure

Congestive Heart Failure (CHF) is a chronic medical condition caused by heart muscle weakness that results in inefficient blood flow in the body. The Swedish Board of Health and Social Welfare defines CHF as the presence of at least one diagnose in Tab 2.2. CHF is a common condition with nearly 26 million people suffering from it worldwide [37]. In Sweden the prevalence of CHF was found to be 2.2% in 2010 [38]. It is also a leading hospitalization cause in Sweden with nearly 44% of CHF patients admitted at least once per year [39]. While the demand of CHF care is understood to be high, there exists limited understanding of how the cost for CHF care is distributed within the healthcare system. The main objective of this study was to leverage the Regional Healthcare Information Platform to understand healthcare utilisation and care patterns of the CHF population while simultaneously exploring potential opportunities of improvement to curb care costs. It gave an experience of conducting retrospective observational studies in medicine. Moreover, from a data science perspective, this study helped in understanding the database content in depth and preparing tools for extracting, merging, customizing, and cleaning the variables with supervision from domain experts. In a typical data science task, these pre-processing steps are often considered to consume nearly 80% of the total effort [40; 41].

ICD 10 code	Description
I110	Hypertensive heart disease with heart failure
I420	Dilated cardiomyopathy
I423	Endomyocardial (eosinophilic) disease
I424	Endocardial fibroelastosis
I425	Other restrictive cardiomyopathy
I426	Alcoholic cardiomyopathy
I427	Cardiomyopathy due to drug and external agent
I428	Other cardiomyopathies
I429	Cardiomyopathy, unspecified
I430	Cardiomyopathy in infectious and parasitic diseases classified elsewhere
I431	Cardiomyopathy in metabolic diseases
I432	Cardiomyopathy in nutritional diseases
I438	Cardiomyopathy in other diseases classified elsewhere
I500	Congestive heart failure
I501	Left ventricular failure
I509	Heart failure, unspecified

Table 2.2: Qualifying ICD-10 codes for congestive heart failure

3. PREDICTIVE MODELLING

In general, prediction modelling has three components: target data, predictor data and a model that maps the relationship between the two. In this chapter we explain how the three axes were constructed in this thesis. Major focus is placed on selection and representation of the predictor data. We present a deep learning framework in which both human and machine derived features are fed sequentially in a cost-sensitive LSTM model to predict patient outcome. To get a flavour of traditional predictive modelling research, we also developed several baseline machine learning models using task-specific expert features. We also demonstrate a simple financial analysis to estimate possible cost-savings if the prediction models are implemented in real clinical workflow. The methodological descriptions are followed by two clinical application tasks: mortality and readmission prediction.

3.1 Ethics

All the studies were conducted using data between 2012 and 2016 from the Regional Healthcare Information Platform with approval from the Ethics Committee in Lund, Sweden (Dnr. 2016/517). Individual informed consent was not requested, but patients were given an opportunity to opt out from participation. For few experimental studies, we have also used the Medical Information Mart for Intensive Care (MIMIC-III) after gaining access rights from PhysioNetWorks [42].

3.2 Methods

A typical prediction model has three axes.

- Target or outcome data: Data about the outcome that we want to predict, for instance, mortality risk.
- Predictor data: Data that is being used to make a prediction, for instance, patient symptoms, age, demographics, clinical history etc.
- ML model: A mathematical function that maps the relationship between the predictor data and the outcome data, for instance, decision trees, neural network etc.

Outcome data, in a retrospective study, is usually extracted directly from EHRs often using a couple of conditional statements. For instance, given the outcome of interest be all-cause *readmission* within 30-days from hospital discharge then *if patient's next admission date - hospital discharge date is <= 30 days; outcome = 1 (readmission) else outcome = 0 (no readmission).* We have defined the outcome measures (mortality and readmission) according to CMS guidelines [43].

While the current outcome measures and definitions and debatable, and we believe that they can be improved by including patient and provider specific experiences [44], we do not consider them in this thesis. However, as patient-reported outcome and experience measures are being developed and validated, new outcomes measures will be part of modern EHRs in the future that are better reflective of the patient's state [45]. In this thesis we primarily focus on selection and representation of the predictor data and training models.

3.2.1 Data representation

To facilitate reading, some key terms are explained below:

- **Clinical profile**: Includes list of clinical codes (or concepts) related to diagnoses, procedures, labs, medications etc. Unless stated otherwise, in this thesis the codes in the clinical profile correspond to a single visit.
- **Visit profile**: This includes the clinical profile along with demographical information like age, gender, visit type etc - Fig 3.1.
- **Patient profile**: Time ordered sequence of visit profiles.
- **Human-derived features (HDF)**: List of features manually selected based on literature review and expert opinion. All demographical features and severity scores are included in this context.
- **Machine-derived features (MDF)**: These are automatically learned features directly from clinical profiles.

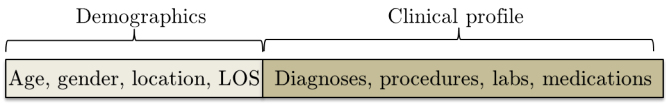

Figure 3.1: A typical visit profile. LOS is the length of stay (in days) of the visit.

A patient in an EHR is often represented as a sequence of care visits. The information in each visit can be broadly categorized into demographics (of patient and care provider) and clinical profile of the patient. Demographic profile includes age, gender, place, type of visit and more. The clinical profile in EHR is represented as a list of clinical codes or concepts pertaining to diagnoses, procedures, lab results, vitals or medication recorded in that visit. Clinical profile can be numerically represented as a scalar severity score, or a feature vector.

As a first step, we have used scalar severity scores to numerically represent the clinical state of the patient during a visit. Severity scores are often computed based on the presence or absence of a limited number of clinical concepts in the patient's history, such as the Charlson Comorbidity Score [46]. While severity scores are easy to calculate and are widely used in clinical decision making today, they scale poorly. Scoring systems are required to be updated with time as new diagnostic, therapeutic and prognostic techniques become available [47]. They are also much dependent on the clinical event, setting and

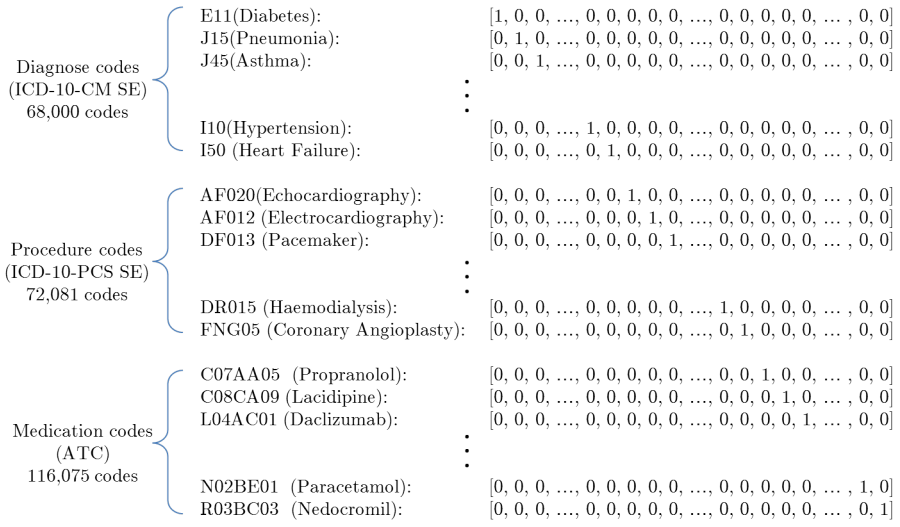

Figure 3.2: One hot encoding. ATC: Anatomical Therapeutic Chemical *Classification System*

application; as misapplication, of such scores can lead to cost and resource wastage [48].

An alternative and straightforward approach is to numerically represent each clinical concept as a binary computable events (one-hot vector) - Fig 3.2. Given N unique clinical codes in EHR, every code is represented as an N dimensional vector with one dimension set to 1 and the rest to zero. The overall clinical profile of the visit can be obtained by summing up all corresponding one-hot vectors. This way clinical representation has two obvious drawbacks.

- It results in high-dimensional and sparse clinical vectors. The dimensionality, to some extent, can be reduced by leveraging the hierarchical structure of the coding schemes and using only higher order (less specific) codes. For instance, in this work, we have considered categorical ICD-10 codes (first three digits) rather than complete 5- 7 digit codes. To further reduce the feature space, we discarded several clinical codes based on their aggregate scores such as low count and variance. This is because much clinical codes have rare occurrences in EHRs.

- It does not account for latent relationships between clinical codes. For instance, hypertension is more closely related to heart failure than ocular pain.

In order to reduce dimensionality while simultaneously preserving latent similarities between clinical concepts, we utilised some algorithms from the Natural Language Processing (NLP) toolbox. NLP constitutes algorithms designed for numerically representing raw texts for computer applications like machine translation, classification, sentiment analysis and more. The NLP toolbox consists of several *word embedding* techniques using dimensionality reduction methods, neural networks (NN), probabilistic models and more. Two promising NN models that attracted a great amount of attention in free text processing were proposed by Mikolov et al [49] - often referred as *word2vec*. These are shallow (one-hidden layer) NNs. Given a *word* in a free text, the goal of word2vec is to predict the neighbouring words - often known as the context - or the other way around: given a context, predict the word. The former is known as Skip Gram (SG) and the latter is the Continuous Bag Of Words (CBOW) model. Intuitively the learning is based on the co-occurrence of words in a similar context. This is referred as *contextual similarity*. For instance, *burger* and *sandwich* are often used with similar context words like *eat*, *cafe*, *cheese*, *lunch* etc. This implies that *burger* and *sandwich* are similar and will have vectors that are close. Put differently, words with similar neighbours are hypothesized to have similar meaning. However, it is an over oversimplifican in the context of languages. Antonyms might also appear in similar contexts. For instance, *tremendous* and *negligible* are also often used with similar context words like *size*, *magnitude*, *range* etc. Though, *tremendous* and *negligible* are not similar words. Despite this limitation, the SG and CBOW models are widely accepted for processing text data because they are still a better alternative to one-hot encodings.

Similar to text data, EHRs possess clinical profiles for every visit which is a list of clinical events. Now the analogy between free text and structured clinical profiles is simple: a clinical profile is considered as a sentence or context and clinical codes as the words in it. And the goal is to construct mathematical vectors for every word (or clinical code) based on its co-occurrences in a given context (or clinical profile). Traditional word embedding tools are built for language texts and consider the order of the words using a fixed size sliding window. Clinical codes in visits, on the other hand, are unordered and each visit may have different numbers of codes. Thus, we slightly modified *word2vec* to support dynamic window size with respect to the size of the clinical profile. Moreover, unlike language text, the notion of 'antonyms' doesn't exist for clinical codes. Learned relationships between ICD-9 codes using *word2vec* on the MIMIC-III dataset can be visualised online [1]

[1] `https://awaisashfaq.com/scatterplot/`

So far, the representations learned from the aforementioned word embedding techniques are those of clinical concepts (individual codes) and not the complete clinical profile. For that task, we leveraged the Paragraph Version of Distributed Bag of Words (PV-DBOW) [50] which is similar to *word2vec* but with an additional paragraph ID (or Visit ID with EHR data) for each profile to store the representation of the complete profile. We also call these machine-derived features (MDF). Finally, a visit representation is created by simply appending MDF and HDF. PV-DBOW training is detailed in Appendix.

A key limitation of NLP inspired embedding techniques is that they do not account for rare words in the vocabulary. Rare clinical codes like cancer or HIV can have significant importance if present in the patient profile. Thus we add more HDF like severity scores (Charlson Comorbididity) to capture the rare events. Other HDF considered in this thesis include the number of prior patient visits to emergency department, inpatient care and outpatient care and medication compliance.

3.2.2 Training models

Given the feature vector generated using HDF, MDF or both, and the target value, the next step in the predictive modelling process is to train a model that learns the mapping function between the feature vector and the target. Among many network architectures, Recurrent Neural Networks (RNNs) have garnered significant attention for predictive modelling of EHRs due to their sequential nature. RNNs belong to the family of artificial neural networks (ANN) with recurrent connections in hidden layer [51; 52]. Operationally, the hidden state h_t is sequentially updated depending on both the activation of the current input x_t at time t, and the previous hidden state of the layer h_{t-1}. Popular RNN variants include the long short-term memory (LSTM) [53] and gated recurrent unit (GRU) models [54]. The need for these variants was triggered by limitations observed when training traditional RNNs because the gradients would often times vanish or (rarely) explode [55]. Thus common gradient-based optimization methods suffered, not only because of varying gradient magnitudes but also because the effect of short-term dependencies would dominate the effect of long-term dependencies (being exponentially smaller with respect to sequence length). In order to address the issue of vanishing gradients for long sequences, two dominant approaches have been explored. First, formulate an alternative training algorithm like clipped gradient as opposed to the stochastic gradient [56]. Second, update the standard RNN units (or cells) to include an internal cell state (in addition to the hidden state) that regulates the flow of information via a set of gates. Thus contrary to traditional RNNs, LSTMs

and GRUs contain an internal recurrence loop at each unit along with three and two gates respectively to control information flow. Gated RNNs have shown to capture long-term dependencies in data and overcome the vanishing gradient problem [57; 58].

A patient in an EHR consists of a sequence of visits and if each visit is represented as a numerical vector (as explained earlier), then those vectors can be sequentially provided as inputs to the RNN model to predict an outcome of interest. In this thesis, we have primarily considered the LSTM network as the prediction model - depicted in Fig 3.3. Given a patient p with T_p visits, the model accepts one visit at each time step. From $t = 2$ till $t = T_p$, new LSTM state is dependent on the state at the previous time step and the new input visit. Of note, at every time step, the LSTM block propagates its state to the next dense layers. This LSTM configuration is often referred to as *sequence to sequence* prediction. This means that at each time step, there is an input and output for the network. We detail the training steps involved in the Appendix.

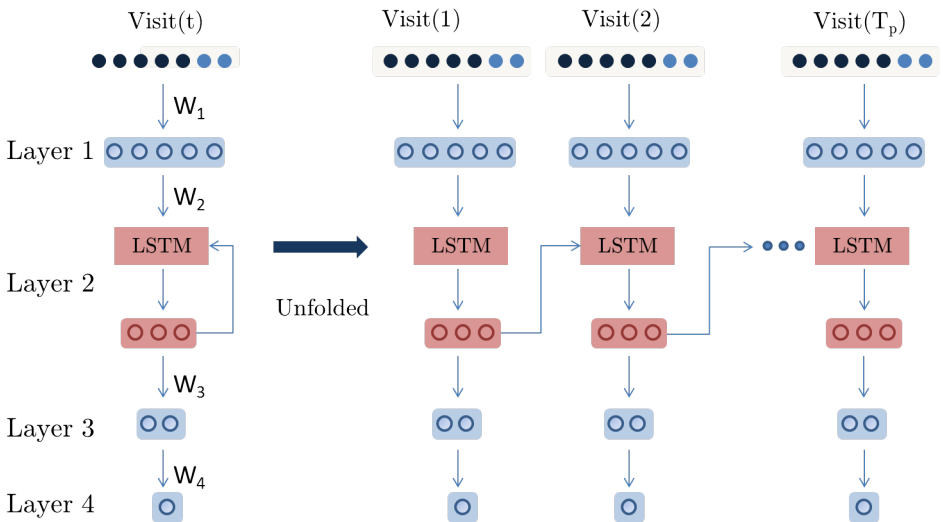

Figure 3.3: Sequential modelling via LSTM. *W*s include the weight matrices and bias terms between layers.

In addition to LSTMs, we also experimented with several baseline prediction models including L2 regularized logistic regression (LR) [59], support vector machine (SVM) [60], K-nearest neighbours classifier (KNN) [61], boosted gradient trees (AB) [62], Random Forest (RF) [63] and Multilayer-Perceptron (MLP) [64].

3.2.3 Data bias correction

In this thesis, so far, we have considered the class imbalance problem (CIP) in the context of *bias*. CIPs are typically addressed from a data and model perspective. A common approach from the data perspective is to oversample or undersample the minority and majority class respectively. This approach has some limitations. Oversampling minority class (or duplicating the instances of the minority class) might result in model over-fitting (i.e. the model function is too closely fit to the minority class instances). Undersampling the majority class might leave out noisy, important or both instances from the data. Thus we might lose important information about the differences between the classes. An alternative approach is the Synthetic Minority Over-sampling Technique (SMOTE) [65]. Instead of oversampling the minority class, SMOTE generates new instances that are interpolations of the minority class and simultaneously under samples the majority class.

In this thesis, we have addressed CIP from a model perspective using the concept of cost-sensitive learning which has shown to outperform sampling techniques in applications where instances are in order of thousands or greater [66]. We begin by defining a cost function $C(A,B)$ that specifies the penalty of misclassifying an instance of class A as class B. $C(A,B)$ is a simple matrix with misclassification cost of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) - Tab 3.1 [67]. c_{TP} and c_{TN} are typically negated costs when the prediction is correct. For c_{FN} and c_{FP} the cost was selected to be equal to the inverse of the proportion of the dataset that the class makes up. $C(A,B)$ is embedded in the learning algorithm of the classifier during training [33]. For instance, let's consider MLP with a softmax output layer and cross-entropy loss function L given as $-\sum_i t_i \log y_i$ where t_i is the target class indicator. y_i is the output of the softmax function. The loss gradient with respect to the input of the sotmax layer z_i is given as $\frac{\partial L}{\partial z_i} = y_i - t_i$. If we add a cost weight α for each class, then the loss function is $-\sum_i \alpha * t_i \log y_i$ and corresponding gradient becomes $\alpha * (y_i - t_i)$. Put differently, if class A has higher weights than class B, then the gradients computed from samples of class A will be greater which in turn will affect weight updates in favour of class A. Similarly, cost-sensitive learning applies to other loss functions and learning algorithms. For example, in RF learning, the class weights are used in two places. First, during the tree induction phase to weight the Gini criterion for finding splits. Second, in the terminal node during aggregation phase where the class prediction is determined via weighted majority votes [68; 69].

	Actual positive y_i	Actual negative y_i
Predicted positive c_i	c_{TP}	c_{FP}
Predicted negative c_i	c_{FN}	c_{FN}

Table 3.1: Typical cost matrix for binary classification

3.3 Economic analysis

An important clinical utility of the prediction models learned in this thesis is to classify if a particular patient visit demands a special intervention due to an increased risk of an adverse outcome. For instance, in readmission prediction application, we predict for each inpatient visit if the patient will be readmitted within 30 days from discharge. Put differently, there exist high readmission-risk and low readmission-risk inpatient visits. In order to the map the output value of a prediction model (like logistic regression, softmax classifiers etc.) into a binary category, we define a decision or discrimination threshold θ. In this thesis, we propose choosing θ by looking at the economic utility of the model. Carrying on with the readmission prediction application, we estimated the potential annual cost savings C_{saved} if an intervention I is selectively offered to patients at high risk of readmission according to Eq. 3.1. Of note, the word *cost* is expressed in monetary terms here.

$$C_{saved} = (C_r \cdot T_r \cdot I_{sr}) - (C_i \cdot P_r) \tag{3.1}$$

where C_r and C_i are the readmission and intervention cost per patient; and I_{sr} is the intervention success rate. T_r and P_r are the number of *truly predicted* and *all predicted* readmissions. The intervention cost is the cost spent on each patient that is predicted with a high readmission risk. Put differently, it includes true positives and false positives.

Since, each visit in the Regional Platform is associated with a cost, C_r in 2016 was approximated to be equal to the mean cost of all readmissions in that year. Now if C_i and I_{sr} are known, T_r and P_r can be calculated over a range of $\theta \in (0,1)$ and θ yielding the maximum C_{saved} can be used to discriminate high and low risk visits. However, in this thesis, we didn't focus on selecting appropriate clinical interventions to prevent adverse outcomes and thus present a spectrum of possible intervention costs, their success rates and corresponding expected savings (Paper III).

3.4 Application and results

The developed models are empirically evaluated in the context of predicting adverse clinical outcomes like mortality and hospital readmissions.

3.4.1 Mortality prediction

The clinical objective of this study was to develop ML models to predict 30-day mortality risk in patients discharged from the Emergency Department (ED). Patients at a higher mortality risk can be considered for end-of-life care planning. The region has two EDs and visits from both EDs between 2015 and 2016 were included in the study. The dataset was highly imbalanced with only 0.21% and 0.15% of the outcome visits belonging to the positive class in the training and testing set respectively. Each visit was assigned a binary outcome conditioned if the patient passed away within 30 days of discharge from the ED. The visits were described using 18 features selected based on relevant literature and guidance from domain experts. Fig 3.4 lists the features used and the correlation matrix. Six different algorithms were selected for training, based on their principally different approaches to prediction. These were LR, SVM, KNN, AB, RF and MLP. To mitigate the class imbalance problem, we over-sampled the minority class in the training set for KNN to equal proportions. For the other algorithms, we used the cost function as explained in section 3.2.3. The models were trained with data from one ED and tested on data from the second ED. All models were optimized for area under the ROC-curve (ROC-AUC). Once the optimal set of hyper-parameters was identified through systematic grid search, the performance of each model was reported on the test set.

SVMs achieved the highest ROC-AUC of 0.95 (95% CI 0.93 - 0.96), followed by LR 0.94 (95% CI 0.93 - 0.95), and RF 0.93 (95% CI 0.92 - 0.95), on the test set. Patient age and co-morbidity score displayed the highest relative importance among the independent variables, followed by arriving in the ED by ambulance. As a continuation of this study we are in the process of conducting a formal cost-benefit analysis, in order to identify associated interventions that are safe, effective and add value to the care system. The study is detailed in Paper II.

3.4.2 Readmission prediction

This study primarily focused on CHF population. Unscheduled readmissions are a hallmark of CHF, with 1 in 4 patients being readmitted within 30 days of discharge [70]. Readmissions are problematic because they pose additional

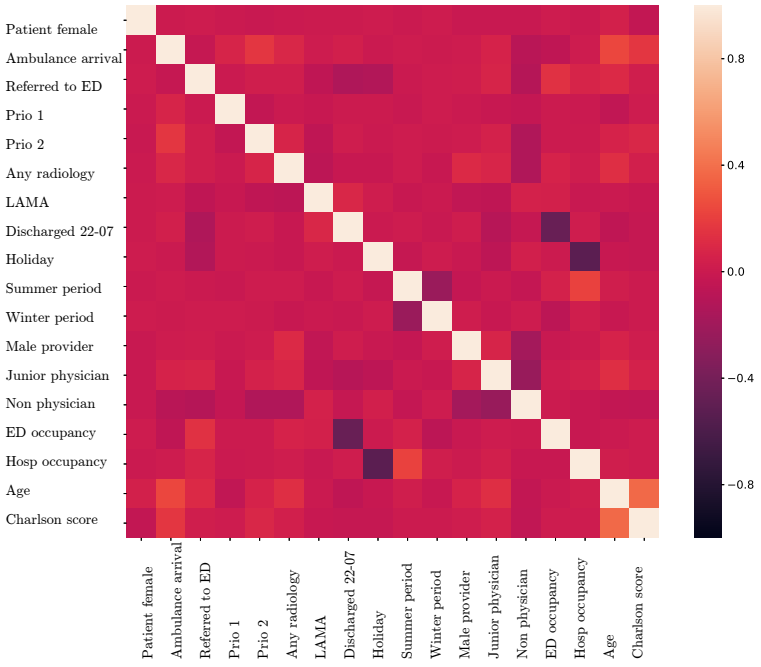

Figure 3.4: Correlation matrix of predictors.

economic burden on the healthcare system and put patients at risk of hospital-acquired infections and clinical errors [71]. They are also considered as a proxy by authorities (like the Centre for Medicare and Medicaid Services in the US) to measure care quality since readmissions are often related to premature discharge or improper treatment [72]. Precise prediction of readmission risk can support care-providers to decide if a patient is ready for discharge or should be considered for an intervention program, eventually reducing the number of unscheduled readmissions and curbing healthcare cost [28].

A recent review investigated 60 studies with 73 models to predict unscheduled 30-day readmissions [73]. It reported moderate discrimination ability. We found that common limitations include building a predictive models that:

- Use either only human-derived features [18; 74] or machine-derived features [75; 76]. The former discards a huge proportion of information in each patient's record, while the latter ignores knowledge and guidelines coming from human intelligence.

- Ignore the sequential or temporal trajectory of events embedded in Electronic Health Records (EHRs) [33; 74; 77–79]. EHRs include a se-

quence of measurements (clinical visits) over time which contains important information about the progression of disease and patient state.

- Fail to consider the skewness in terms of class imbalance and different costs of misclassification errors [74; 75; 78–81]. Class imbalance problems are common with EHR data [82]. A favourable prediction model is often the one with a high precision on the minority class with a reasonable precision on the majority class.

Though several studies, to some extent, have independently addressed one (or sometimes two) of the aforementioned limitations; to the best of our knowledge, there is no single model that addresses all the three limitations. The overall framework in this study is illustrated in Fig. 3.5. Each visit was represented by 9 HDF and 185 dimensional MDF vector generated via the PV-DBOW model. Tab 3.2 shows the list of HDF used to describe a visit in this study. The LSTM network is depicted in Fig 3.3. The data has a class im-

Age at the time of visit	Discrete
Gender	Binary
Medication compliance	Binary
Duration of stay	Discrete
Type of visit	Binary
Charlson comorbidity score	Discrete
Number of prior emergency care visits	Discrete
Number of prior admissions	Discrete
Number of prior outpatient visits	Discrete

Table 3.2: Human-derived features for each visit. The feature 'compliance' reports medicine prescription of both ACE inhibitors and Beta Blockers for CHF patients. It is based on national guidelines by Socialstrylesen in Sweden. Types of visit include scheduled and unscheduled admissions. The prior visits were counted on a 6-month window starting from the date of admission.

balance ratio of 0.28, thus, a cost function was embedded in the loss function during training. We conducted an iterative design of 12 experiments with possible combinations of model characteristics that are summarized in Tab 3.3.

We split the data into training and test sets in two different ways. In case 1, we split based on patients. 70% patients were used for training and 30% for testing and we report the performance on the test set. In case 2, we split based on time. We used all patient data from 2012-2015 for training and tested on the complete set 2012-2016. However, we evaluate model performance

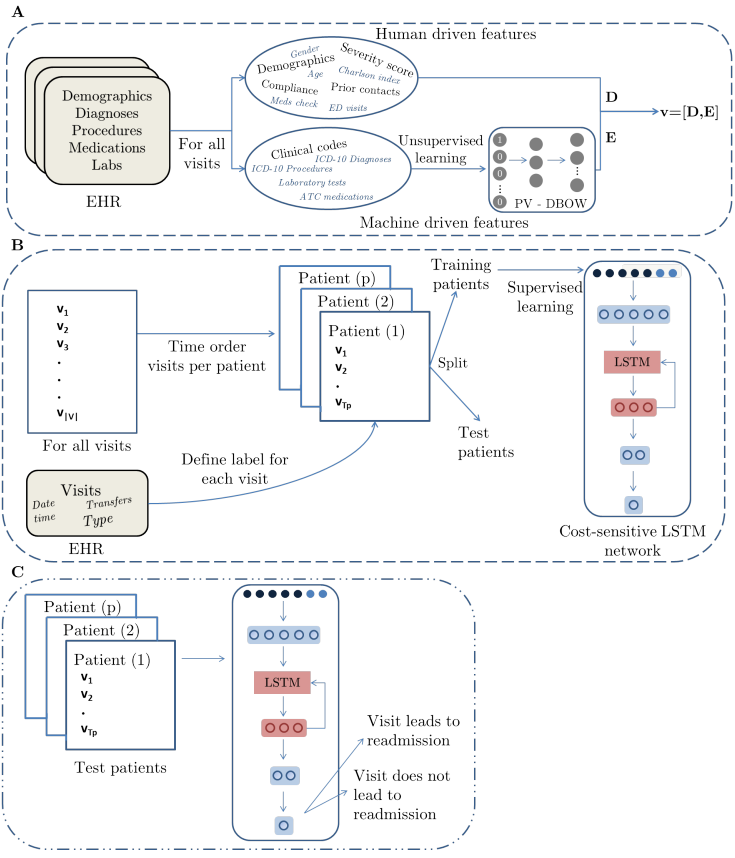

Figure 3.5: Proposed framework. A: Generate visit representations from human and machine-derived features. The output is a feature vector for each visit. B: The visit representations are fed sequentially (per patient) in a cost-sensitive LSTM network for training. C: The test patients are fed to the trained network to predict 30-day readmission risk at each visit.

Characteristic	Meaning
HDF	Human derived features are fed as input to the model
MDF	Machine derived contextual embeddings are fed as input to the model
LSTM	The model capture the sequential visit patterns in the EHR
CA	The model adjusts for misclassification costs

Table 3.3: Model characteristics explained

on visits that occurred in 2016 only. In the first case, we make sure that no patient overlaps between training and test set. Thus, the model is tested on new patients that weren't part of training. This is the usual way of evaluating readmission predictions in previous studies. However, practically if a model is

used in a clinical work, it would require predicting readmission risk for new patients (with no prior admission) and patients with prior admissions. The model performance on future visits by the latter patients can be improved if we include their prior data in the training set. Thus we see better performance if data is split according to case 2.

We show that leveraging both human and machine-derived features from EHR, together with an LSTM network outperforms models that ignore any of these characteristics - Tab 3.3. For both case 1 and 2, the model with all four characteristic reached an AUC of 0.77 (SD 0.006) and 0.82 (SD 0.003) and marked the best value. Fig 3.6 shows the results for case 1. As a next stage of the project, we aim to explore attention-based learning on LSTMs to provide model intelligibility. Knowing what features contributed to a certain outcome can facilitate designing relevant interventions to prevent or prepare for adverse outcomes. The complete study is presented in Paper III.

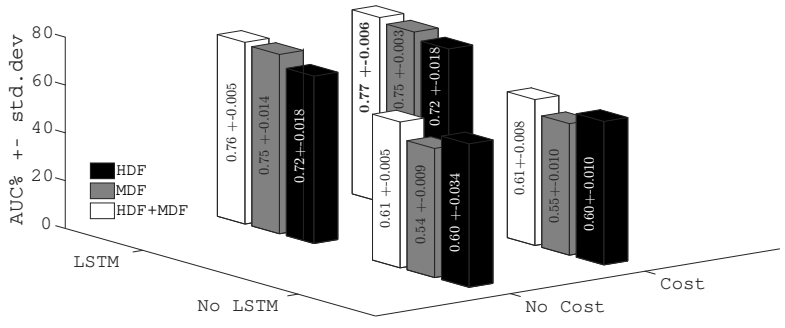

Figure 3.6: Model performance on predicting 30-day readmission on test patients. The model was trained on 70% of CHF patients and tested on the rest.

4. CONCLUDING REMARKS

In this thesis, we have presented methods to represent sequential patient information in EHRs in order to predict adverse clinical outcomes. The key challenges of heterogeneity and temporality are addressed using human-derived features appended to contextual embeddings of clinical concepts and LSTMs respectively. From a clinical perspective, patients at risk of an adverse outcome can facilitate from personalized intervention programs designed to eliminate or prepare for them. From an economic perspective, eliminating or preparing for adverse outcomes (like unscheduled readmissions) means huge reduction in healthcare costs and precise resource utilization. From a data science perspective, the thesis demonstrates how different ML models (primarily designed for other application domains) can be combined and adapted to address EHR related modelling challenges.

Though insights into future (like risk of readmission) allow informed decision-making, in order to address the root cause of readmission and select effective interventions, it is important to understand what group of features contributed to the prediction. Immediate future work will focus on models that facilitate interpretation of LSTMs and contextual embeddings. In an effort to do so, this chapter presents a brief overview of an ongoing study and future directions.

4.1 Ongoing work and future directions

Understanding the reason behind a model's prediction is particularly important in critical applications like healthcare. Though there is no real consensus to define and assess intelligibility of a ML model, as of now, we define *model intelligibility* as quantifying the contribution/importance of input variables towards the prediction score. In an effort to do so, we already began with a RF model using multi-hot representations of complete clinical profile of a patient visit to predict several adverse clinical outcomes. The study is centred on each hospital admission by the CHF cohort. The observation window to include clinical records is set to 365 days prior to discharge date. In order to capture the sequential nature of EHRs, in each visit representation, we define three historical features that count the number of outpatient (specialty + primary care), emergency and inpatient visits prior to that visit. A typical visit is represented as shown in Fig 4.1. In total, 65,111 hospital admissions were considered, out of which 70% were used for training and 30% for testing. Preliminary results on the test set are reported in Tab 4.1. Important predictors of mortality include lung and kidney disorders, length of stay (LOS) and age. For readmission prediction, the most important predictors were the prior visit counts followed by hospitalization days, age and visit location.

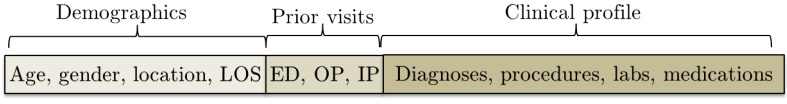

Figure 4.1: Inpatient visit representation. LOS is measured in days for each visit. The clinical profile is represented as a 1256 dimensional multi-hot vector.

Other preliminary findings (unrelated to intelligibility) from this study have led us to think over the following research directions in future:

- For readmission prediction, RF performed poorly compared to the LSTM network explained in the section 3.4.2 in terms of ROC-AUC. Thus, to have both high performance and intelligibility, methods to improve RF performance or interpret LSTMs as well the contextual embeddings need to be explored and developed.

- In this study, we found lab exams to be among the least relevant features for prediction. A possible reason might be that we consider labs as binary indicators (either the patient took the test or not) rather than recording true lab results. Several recent studies tend to ignore true lab

results for EHR driven prediction tasks because their automatic integration to a deep learning model is challenging [31; 76; 83]. However, lab results, in general, are considered important predictors of patient outcomes [84; 85] and we aim to develop a strategy to encode true lab results within the visit representations.

- Last but not least, investigating the impact of important predictors, we notice variation in the mean ROC-AUC when the model is tested on a subset of the test set - Tab 4.1. The differences are prominent in subsets centred over age and number of previous admissions. Apart from the RF model, variations in mean ROC-AUC are also observed using the cost-sensitive LSTM model (Tab 4.2) explained in section 3.4.2. The rationale for such variations needs to be explored and, if necessary, precise confidence intervals should be determined around individual predictions. This is because in clinical setting, the model predictions may influence crucial clinical decisions concerning the course of action for improving patient's health. The concept of having error bounds on a per-instance basis is often referred as *conformal prediction* but its application on EHR driven prediction models has been scarce [86].

	Outcome		
Test set	In hospital mortality	30- day mortality	30- day readmission
Complete	0.87 (0.001)	0.79 (0.010)	0.61 (0.001)
Male	0.86 (0.001)	0.83 (0.012)	0.59 (0.001)
Female	0.88 (0.001)	0.79 (0.015)	0.60 (0.000)
Age ≥ 60	0.87 (0.001)	0.80 (0.010)	0.59 (0.000)
Age < 60	0.95 (0.002)	0.88 (0.055)	0.64 (0.002)
Hospital A	0.87 (0.001)	0.83 (0.001)	0.59 (0.000)
Hospital B	0.88 (0.002)	0.79 (0.016)	0.60 (0.001)
In count ≥ 3	0.85 (0.001)	0.78 (0.012)	0.61 (0.000)
In count < 3	0.92 (0.001)	0.86 (0.010)	0.55 (0.001)

Table 4.1: Preliminary results of the RF model on the complete test set and its subsets: Reporting ROC AUC and std. deviations. 'In count' specifies the number of hospital admissions or inpatient visits by the patient.

	Outcome
Test set	30- day readmission
Complete	0.77 (0.006)
Male	0.75 (0.045)
Female	0.72 (0.049)
Age ≥ 60	0.76 (0.006)
Age < 60	0.78 (0.028)
In count >3	0.77 (0.008)
In count < 3	0.74(0.006)
LOS$\geq$3	0.77 (0.007)
LOS$<$3	0.75 (0.007)
Charlson score $\geq$5	0.79 (0.013)
Charlson score < 5	0.75 (0.006)

Table 4.2: Preliminary results of the LSTM model on the complete test set and its subsets: Reporting ROC AUC and std. deviations. 'In count' specifies the number of hospital admissions or inpatient visits by the patient.

4.2 Conclusion

In this thesis, we have presented methods that, to a large extent, address the heterogeneity, temporality and class imbalance challenges that impede effective learning EHR driven prediction models. To that end, we have leveraged knowledge created from three mature research fields: language processing, recurrent neural networks and cost-sensitive learning. Moreover, the contribution from clinical science like different hierarchical classification schemata and expert features cannot be ignored. One value of this thesis is a machine learning frame work in which both human and machine derived features (that together capture a broad picture of the patient's health state) are fed in a cost-sensitive sequential model to predict adverse outcomes. This, in turn, can influence clinician's decisions (for admissions, transfers, discharge, interventions, etc.) for better. However, there is more to it.

While the developed models can influence clinical decisions for good, they do not help us solve the root cause of adverse outcomes. This is because we are often unaware of what variables or groups of variables most effect the prediction score and how. Of note, it is critical to understand if the effects of variables on the prediction score are true causations or mere correlations. This is because correlation between two variables does not guarantee if one is the *cause* or *result* of the other [87]. The input variables to the prediction model can be grouped into two types: non-modifiable (like age, gender, diagnoses

etc.) and modifiable (like medications, procedures, care units etc.). Though all variables contributing causally to prediction score are of interest, from an administrative perspective, highlighting modifiable variables are of prime significance as it allows decision-makers to radically investigate the utility of the resource and recommend changes in care processes if necessary. Put differently, the goal is not limited to accurate prediction of adverse outcomes but also to facilitate restructuring of care delivery in a way that reduces the number of cases for adverse outcome in future. Thus reiterating the clinical research question presented in this thesis - *What is the probability of a given outcome for a patient at a specific time given data available at that time?* - the next step is to answer: *What decides the outcome of the patient?* From a ML perspective, the next question in hand is: *How can black box models like LSTMs be modified to quantify the importance of input features to the prediction score?*

Last but not least, a significant portion of the thesis work, so far, has been focused on exploring the complexity surrounding EHR analysis and developing a comprehensive library (with user-guides and functions) to access, clean and manipulate entire EHR data. This was made possible by collaborating with doctors, administrators, health economists and database engineers on multiple projects. The developed library has been used to conduct both retrospective observational studies and predictive modelling. Though not particularly in the context of *research*, building the library in itself carries a significant internal value since it was clearly lacking in this particular field. The library can now serve as the foundation to build new tools that can have meaningful research and economic impact.

Appendix

Learning visit representations

Fig 4.2 shows the paragraph-version of distributed bag of words (PV-DBOW) architecture. Given $|V|$ and $|S|$ unique visit IDs and clinical codes in the data respectively, each visit ID and clinical code is represented as an independent $|V|$ and $|S|$ dimensional vector $\boldsymbol{v'_n} \in \{0,1\}^{|V|}$ and $\boldsymbol{s'_m} \in \{0,1\}^{|S|}$ with one index set to 1 and rest to 0. Consider a visit profile like fig 4.3 with 10 clinical codes $C_v = \{c_{v1}, c_{v2}, ..., c_{v10}\}$ against a visit ID v'_n. Following the PV-DBOW model, the input $\rightarrow$ target pairs ($\boldsymbol{x} \rightarrow \boldsymbol{t}$) for training are $\boldsymbol{v'_n} \rightarrow \boldsymbol{s'_1}$; $\boldsymbol{v'_n} \rightarrow \boldsymbol{s'_2}$ and so on till $\boldsymbol{v'_n} \rightarrow \boldsymbol{s'_{10}}$. $\boldsymbol{t}$ is a one-hot target vector also of size S. The number of nodes K in the hidden layer h is a hyper-parameter that represent the mapping size of the Visit ID. The activation function of h is linear and thus its output H is matrix multiplication of input $\boldsymbol{v'}$ and weight matrix E. Output layer has a softmax activation function. The output y_j is given as

$$y_j = \frac{exp^{z_j}}{\sum_{m \in K} exp^{(z_m)}} \tag{4.1}$$

where $\boldsymbol{z} = H.E'$. Softmax normalizes the exponential of $\boldsymbol{y}$. Put differently, given a Visit ID v'_n as input, every neuron in the output layer y_j represents the conditional probability of corresponding clinical code being in the profile of v'. Given the weight matrices E,E', the training objective then is to maximize $y_{j'}$ where j' is the index of the true clinical codes. However, since we have $|C_v|$ codes for v'_n; the training objective is to maximize $|C_v|$ true target neurons in the output layer $\boldsymbol{y}$. The corresponding cost function J is the negative log probability of all the actual codes - also known as cross entropy. Once trained, each row $\boldsymbol{E_v}$ of the embedding matrix is the numerical representation of the clinical state of visit v. Mathematically,

$$\underset{\boldsymbol{E},\boldsymbol{E'}}{\arg\max} \prod_{v \in V} p(C_v|v) \tag{4.2}$$

The size of representation K is a hyper-parameter and was set to 185 in this work. The optimization was carried out Gensim 3.4.0 with negative sampling in Python [88]. The window size was set to the maximum number of codes in a visit 121.

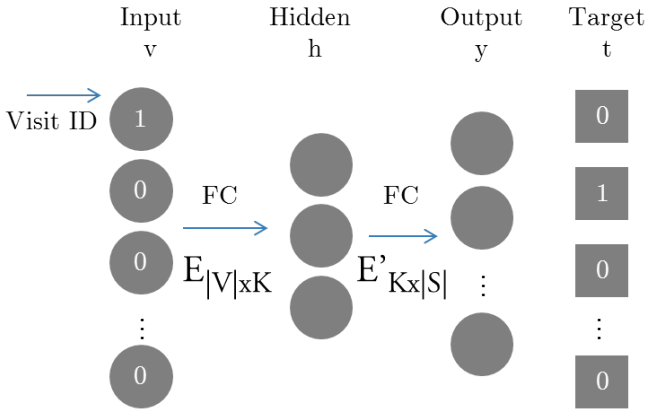

Figure 4.2: Doc2vec: Distributed Bag of Words version of Paragraph Vector (PV-DBOW). FC: Fully connected

C_v I44, I48, I25, AF02, DR02, C07, C09, /HB, /Amylas, /Albu

Figure 4.3: Example: clinical state of a visit

LSTM architecture

The LSTM network is depicted in Fig 3.3. Let P denote a set of all patients where $|P|$ is the total number of patients in the dataset. Given a patient p or a sequence of clinical visits of length T_p, the model accepts each visit as a $K = 194$ dimensional vector $\boldsymbol{v_{pt}}$ (185 from the embedding matrix and 11 expert-features). Layers 1, 3 and 4 with output ($\boldsymbol{h_1}$), ($\boldsymbol{h_3}$) and ($\boldsymbol{h_4}$) are fully connected dense layers with 128, 32 and 1 node respectively and sigmoid activation functions. The size of the hidden layer of LSTM (layer 2) is 64. The input to the LSTM layer ($\boldsymbol{h_2^{in}}$) is $\sigma(W_2\boldsymbol{h_1} + \boldsymbol{b_2})$ where $h_1 = \sigma(W_1\boldsymbol{v_t} + \boldsymbol{b_1})$. $\sigma()$ is the sigmoid function and W and b are the weight matrices and bias terms. Below we expand on the LSTM block [1] in the model Fig 4.4.

$$
\begin{aligned}
\boldsymbol{f_t} &= \sigma(W_{fx}\boldsymbol{x_t} + W_{fh}\boldsymbol{h_{t-1}} + \boldsymbol{b_f}) \\
\boldsymbol{i_t} &= \sigma(W_{ix}\boldsymbol{x_t} + W_{ih}\boldsymbol{h_{t-1}} + \boldsymbol{b_i}) \\
\boldsymbol{o_t} &= \sigma(W_{ox}\boldsymbol{x_t} + W_{oh}\boldsymbol{h_{t-1}} + \boldsymbol{b_o}) \\
\boldsymbol{c'_t} &= \tanh(W_{cx}\boldsymbol{x_t} + W_{ch}\boldsymbol{h_{t-1}} + \boldsymbol{b_c}) \\
\boldsymbol{c_t} &= \boldsymbol{f_t} \cdot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \cdot \boldsymbol{c'_t} \\
\boldsymbol{h_t} &= \boldsymbol{o_t} \cdot \tanh(\boldsymbol{c_t})
\end{aligned}
\tag{4.3}
$$

[1] Image from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

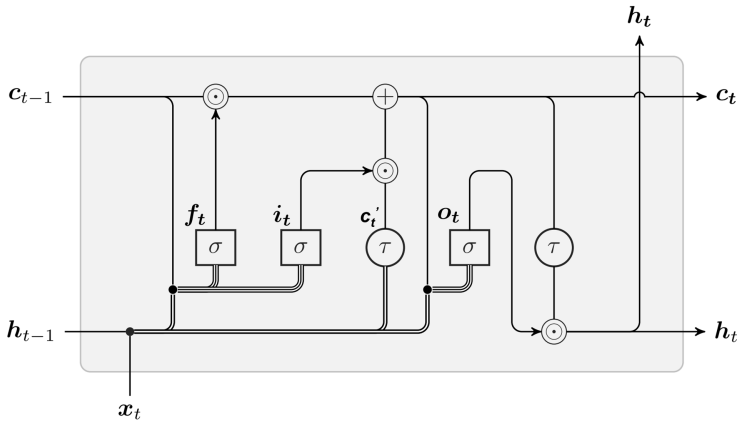

Figure 4.4: LSTM block. Of note $\boldsymbol{x_t}$ $\boldsymbol{h_t}$ corresponds to $\boldsymbol{h_2^{in}}$ and $\boldsymbol{h_2}$ in this work

$\boldsymbol{f_t}$, $\boldsymbol{i_t}$ $\boldsymbol{o_t}$ are referred to as forget, input and out gate respectively that drop or retain the relevant information from previous states [53]. Unlike conventional RNNs where the new hidden state $\boldsymbol{h_t}$ depends only on the input $\boldsymbol{x_t}$ and $\boldsymbol{h_{t-1}}$; the hidden state in LSTM is dependent on gate states and an intermediate memory unit $\boldsymbol{c_t'}$ according to Eq. 4.3. The LSTM block updates its state $\boldsymbol{h_t}$ at every time step t and also propagates it to following dense layers. The final output of the model $\boldsymbol{y_t} = \sigma(W_4\boldsymbol{h_3} + \boldsymbol{b_4})$ is generated at every time step where $\boldsymbol{h_3} = \sigma(W_3\boldsymbol{h_2} + \boldsymbol{b_3})$. Next, the error E (binary cross entropy in our case) is calculated using the model output $\boldsymbol{y_t}$ and true targets $\hat{\boldsymbol{y}}_t$: $E = \sum_i^{C=2} \hat{\boldsymbol{y}}_t log(\boldsymbol{y_t})$. To account for class imbalance problem, a cost term is added. We set $c_{FN} = 3$ and $c_{FP} = 1$ in our experiements with cost adjustments (Table 3.1). For without cost experiments, $c_{FN} = c_{FN} = 1$. To learn and optimize the parameters of the model, we set the binary cross entropy as the loss function and minimize with respect to weights and bias terms W.

$$\min_{\boldsymbol{W}} \sum_{p=1}^{|P|} \sum_{t=1}^{T_p} [-y'_{pt} log(y_{pt}) - (1 - y'_{pt}) log(1 - y_{pt})].C(y'_{pt}) \tag{4.4}$$

where y'_{pt} is the readmission indicator for the t^{th} visit of p^{th} patient where 1 indicates readmission and 0 control. The loss minimization and parameter (W) optimization was performed through back-propagation using mini-batch gradient descent (Batch size = 32 and Max epochs = 80) implemented via Keras 2.2.2 [89].

References

[1] PETER DENSEN. **Challenges and opportunities facing medical education**. *Transactions of the American Clinical and Climatological Association*, **122**:48, 2011. 2

[2] WORLD HEALTH ORGANIZATION ET AL. **The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research**. 1993. 2

[3] WORLD HEALTH ORGANIZATION. *World report on ageing and health*. World Health Organization, 2015. 2

[4] REZA MIRNEZAMI, JEREMY NICHOLSON, AND ARA DARZI. **Preparing for precision medicine**. *New England Journal of Medicine*, **366**(6):489–491, 2012. 2

[5] ZIAD OBERMEYER AND THOMAS H LEE. **Lost in thought—the limits of the human mind and the future of medicine**. *New England Journal of Medicine*, **377**(13):1209–1211, 2017. 2

[6] JOHN NS MATTHEWS. *Introduction to randomized controlled clinical trials*. Chapman and Hall/CRC, 2006. 2

[7] JAROSŁAW WASILEWSKI, LECH POLOŃSKI, ANDRZEJ LEKSTON, TADEUSZ OSADNIK, RAFAŁ REGUŁA, KAMIL BUJAK, AND ANNA KUREK. **Who is eligible for randomized trials? A comparison between the exclusion criteria defined by the ISCHEMIA trial and 3102 real-world patients with stable coronary artery disease undergoing stent implantation in a single cardiology center**. *Trials*, **16**(1):411, 2015. 2

[8] KAREN E JOYNT AND ASHISH K JHA. **Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program**. *Jama*, **309**(4):342–343, 2013. 3

[9] VIKAS WADHWA AND MORVEN DUNCAN. **Strategies to avoid unnecessary emergency admissions**. *BMJ*, **362**:k3105, 2018. 3

[10] RASMUS ÅHMAN, PONTUS FORSBERG SIVERHALL, JOHAN SNYGG, MATS FREDRIKSON, GUNNAR ENLUND, KARIN BJÖRNSTRÖM, AND MICHELLE S CHEW. **Determinants of mortality after hip fracture surgery in Sweden: a registry-based retrospective cohort study**. *Scientific reports*, **8**(1):15695, 2018. 3

[11] ERIN P BALOGH, BRYAN T MILLER, JOHN R BALL, ENGINEERING NATIONAL ACADEMIES OF SCIENCES, MEDICINE, ET AL. **Overview of Diagnostic Error in Health Care**. 2015. 3

[12] NASSER M NASRABADI. **Pattern recognition and machine learning**. *Journal of electronic imaging*, **16**(4):049901, 2007. 3

[13] GEOFFREY HINTON. **Deep learning—a technology with the potential to transform health care**. *Jama*, **320**(11):1101–1102, 2018. 3, 4

[14] BRADFORD H GRAY, THOMAS BOWDEN, IB JOHANSEN, AND SABINE KOCH. **Electronic health records: an international perspective on" meaningful use".** *Issue Brief (commonwealth fund)*, **28**:1–18, 2011. 3

[15] D ADAMSKI. **Overview of the national laws on electronic health records in the EU Member States. National Report for Poland**, 2014. 3

[16] AMY COMPTON-PHILLIPS. **Care Redesign Survey: What Data Can Really Do for Health Care**. *NEJM Catalyst Insights Report*, 2017. 3

[17] CATALINA MARTÍNEZ-COSTA, DIPAK KALRA, AND STEFAN SCHULZ. **Improving EHR semantic interoperability: future vision and challenges.** In *MIE*, pages 589–593, 2014. 3

[18] BENJAMIN A GOLDSTEIN, ANN MARIE NAVAR, MICHAEL J PENCINA, AND JOHN IOANNIDIS. **Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review**. *Journal of the American Medical Informatics Association*, **24**(1):198–208, 2017. 4, 27

[19] YANN LECUN, YOSHUA BENGIO, AND GEOFFREY HINTON. **Deep learning**. *nature*, **521**(7553):436, 2015. 4

[20] RICHARD BELLMAN. *Dynamic programming*. Courier Corporation, 2013. 5

[21] IRA ASSENT. **Clustering high dimensional data**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(4):340–350, 2012. 5

[22] NITESH V CHAWLA, NATHALIE JAPKOWICZ, AND ALEKSANDER KOTCZ. **Special issue on learning from imbalanced data sets**. *ACM Sigkdd Explorations Newsletter*, **6**(1):1–6, 2004. 6

[23] **Special report: The making of an HIV catastrophe**. *DAWN news*, 2017. 6

[24] DENIS AGNIEL, ISAAC S KOHANE, AND GRIFFIN M WEBER. **Biases in electronic health record data due to processes within the healthcare system: retrospective observational study**. *Bmj*, **361**:k1479, 2018. 6

[25] MILENA A GIANFRANCESCO, SUZANNE TAMANG, JINOOS YAZDANY, AND GABRIELA SCHMAJUK. **Potential biases in machine learning algorithms using electronic health record data**. *JAMA internal medicine*, **178**(11):1544–1547, 2018.

[26] ROBERT A VERHEIJ, VASA CURCIN, BRENDAN C DELANEY, AND MARK M MCGILCHRIST. **Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse**. *Journal of medical Internet research*, **20**(5), 2018. 6

[27] HARLAN M KRUMHOLZ, ANGELA R MERRILL, ERIC M SCHONE, GEOFFREY C SCHREINER, JERSEY CHEN, ELIZABETH H BRADLEY, YUN WANG, YONGFEI WANG, ZHENQIU LIN, BARRY M STRAUBE, ET AL. **Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission**. *Circulation: Cardiovascular Quality and Outcomes*, pages CIRCOUTCOMES–109, 2009. 7

[28] SUNIL KRIPALANI, CECELIA N THEOBALD, BETH ANCTIL, AND EDUARD E VASILEVSKIS. **Reducing hospital readmission rates: current strategies and future directions**. *Annual review of medicine*, **65**:471–485, 2014. 7, 27

[29] NATIONAL RESEARCH COUNCIL ET AL. *Scientific research in education*. National Academies Press, 2002. 8

[30] MENNO MOSTERT, ANNELIEN L BREDENOORD, MONIQUE CIH BIESAART, AND JOHANNES JM VAN DELDEN. **Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach**. *European Journal of Human Genetics*, **24**(7):956, 2016. 12

[31] EDWARD CHOI, ANDY SCHUETZ, WALTER F STEWART, AND JIMENG SUN. **Using recurrent neural network models for early detection of heart failure onset**. *Journal of the American Medical Informatics Association*, **24**(2):361–370, 2016. 12, 33

[32] ZACHARY C LIPTON, DAVID C KALE, CHARLES ELKAN, AND RANDALL WETZEL. **Learning to diagnose with LSTM recurrent neural networks**. *arXiv preprint arXiv:1511.03677*, 2015. 12

[33] HAISHUAI WANG, ZHICHENG CUI, YIXIN CHEN, MICHAEL AVIDAN, ARBI BEN ABDALLAH, AND ALEXANDER KRONZER. **Predicting Hospital Readmission via Cost-sensitive Deep Learning**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018. 12, 24, 27

[34] AFEEZAT OLAJUMOKE OYEYEMI AND PHILIP SCOTT. **Interoperability in health and social care: organizational issues are the biggest challenge**. *Journal of innovation in health informatics*, **25**(3):196–198, 2018. 14

[35] MIRIAM REISMAN. **EHRs: The Challenge of Making Electronic Data Usable and Interoperable**. *Pharmacy and Therapeutics*, **42**(9):572, 2017. 14

[36] ERIC J LAMMERS, JULIA ADLER-MILSTEIN, AND KEITH E KOCHER. **Does health information exchange reduce redundant imaging? Evidence from emergency departments**. *Medical care*, **52**(3):227–234, 2014. 14

[37] PIOTR PONIKOWSKI, STEFAN D ANKER, KHALID F ALHABIB, MARTIN R COWIE, THOMAS L FORCE, SHENGSHOU HU, TINY JAARSMA, HENRY KRUM, VISHAL RASTOGI, LUIS E ROHDE, ET AL. **Heart failure: preventing disease and death worldwide**. *ESC Heart Failure*, **1**(1):4–25, 2014. 15

[38] RAMIN ZARRINKOUB, BJÖRN WETTERMARK, PER WÄNDELL, MÄRIT MEJHERT, ROBERT SZULKIN, GUNNAR LJUNGGREN, AND THOMAS KAHAN. **The epidemiology of heart failure, based on data for 2.1 million inhabitants in Sweden**. *European journal of heart failure*, **15**(9):995–1002, 2013. 15

[39] ALDO P MAGGIONI, ULF DAHLSTRÖM, GERASIMOS FILIPPATOS, OVIDIU CHIONCEL, MARISA CRESPO LEIRO, JAROSLAW DROZDZ, FRIEDRICH FRUHWALD, LARS GULLESTAD, DAMIEN LOGEART, GIANNA FABBRI, ET AL. **EURObservational Research Programme: regional differences and 1-year follow-up results of the Heart Failure Pilot Survey (ESC-HF Pilot)**. *European journal of heart failure*, **15**(7):808–817, 2013. 15

[40] GIL PRESS. **Cleaning big data: Most time-consuming, least enjoyable data science task, survey says**. *Forbes, March*, **23**, 2016. 15

[41] STEVE LOHR. **For big-data scientists,'janitor work'is key hurdle to insights**. *New York Times*, **17**, 2014. 15

[42] ALISTAIR EW JOHNSON, TOM J POLLARD, LU SHEN, H LEHMAN LI-WEI, MENGLING FENG, MOHAMMAD GHASSEMI, BENJAMIN MOODY, PETER SZOLOVITS, LEO ANTHONY CELI, AND ROGER G MARK. **MIMIC-III, a freely accessible critical care database**. *Scientific data*, **3**:160035, 2016. 18

[43] JN GRADY, Z LIN, Y WANG, C NWOSU, M KEENAN, K BHAT, H KRUMHOLZ, AND S BERNHEIM. **measures updates and specifications: Acute myocardial infarction, heart failure, and pneumonia 30-day risk-standardized mortality measure (version 7.0)**. *Yale University/Yale-New Haven Hospital-Center for Outcomes Research & Evaluation (Yale-CORE): Technical Report. Accessed August*, **8**:2015, 2013. 18

[44] NICK BLACK. **Patient reported outcome measures could help transform healthcare**. *Bmj*, **346**:f167, 2013. 18

[45] NW WAGLE. **Implementing patient-reported outcome measures**. *NEJM Catalyst. November*, **17**, 2016. 18

[46] MARY E CHARLSON, PETER POMPEI, KATHY L ALES, AND C RONALD MACKENZIE. **A new method of classifying prognostic comorbidity in longitudinal studies: development and validation**. *Journal of chronic diseases*, **40**(5):373–383, 1987. 19

[47] JEAN-LOUIS VINCENT AND RUI MORENO. **Clinical review: scoring systems in the critically ill**. *Critical care*, **14**(2):207, 2010. 19

[48] AMY GRACE RAPSANG AND DEVAJIT C SHYAM. **Scoring systems in the intensive care unit: a compendium**. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, **18**(4):220, 2014. 20

[49] TOMAS MIKOLOV, KAI CHEN, GREG CORRADO, AND JEFFREY DEAN. **Efficient estimation of word representations in vector space**. *arXiv preprint arXiv:1301.3781*, 2013. 21

[50] QUOC LE AND TOMAS MIKOLOV. **Distributed representations of sentences and documents**. In *International Conference on Machine Learning*, pages 1188–1196, 2014. 22

[51] ROBERT J SCHALKOFF. *Artificial neural networks*, **1**. McGraw-Hill New York, 1997. 22

[52] CAO XIAO, EDWARD CHOI, AND JIMENG SUN. **Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review**. *Journal of the American Medical Informatics Association*, 2018. 22

[53] SEPP HOCHREITER AND JÜRGEN SCHMIDHUBER. **Long short-term memory**. *Neural computation*, **9**(8):1735–1780, 1997. 22, 39

[54] KYUNGHYUN CHO, BART VAN MERRIËNBOER, DZMITRY BAHDANAU, AND YOSHUA BENGIO. **On the properties of neural machine translation: Encoder-decoder approaches**. *arXiv preprint arXiv:1409.1259*, 2014. 22

[55] YOSHUA BENGIO, PATRICE SIMARD, AND PAOLO FRASCONI. **Learning long-term dependencies with gradient descent is difficult**. *IEEE transactions on neural networks*, **5**(2):157–166, 1994. 22

[56] YOSHUA BENGIO, NICOLAS BOULANGER-LEWANDOWSKI, AND RAZVAN PASCANU. **Advances in optimizing recurrent networks**. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013. 22

[57] SANDEEP AYYAR, OB DON, AND W IV. **Tagging Patient Notes with ICD-9 Codes**. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, 2016. 23

[58] JUNYOUNG CHUNG, CAGLAR GULCEHRE, KYUNGHYUN CHO, AND YOSHUA BENGIO. **Empirical evaluation of gated recurrent neural networks on sequence modeling**. *arXiv preprint arXiv:1412.3555*, 2014. 23

[59] STROTHER H WALKER AND DAVID B DUNCAN. **Estimation of the probability of an event as a function of several independent variables**. *Biometrika*, **54**(1-2):167–179, 1967. 23

[60] JOHAN AK SUYKENS AND JOOS VANDEWALLE. **Least squares support vector machine classifiers**. *Neural processing letters*, **9**(3):293–300, 1999. 23

[61] THOMAS COVER AND PETER HART. **Nearest neighbor pattern classification**. *IEEE transactions on information theory*, **13**(1):21–27, 1967. 23

[62] YOAV FREUND AND ROBERT E SCHAPIRE. **A decision-theoretic generalization of on-line learning and an application to boosting**. *Journal of computer and system sciences*, **55**(1):119–139, 1997. 23

[63] LEO BREIMAN. **Random forests**. *Machine learning*, **45**(1):5–32, 2001. 23

[64] DAVID E RUMELHART, GEOFFREY E HINTON, AND RONALD J WILLIAMS. **Learning representations by back-propagating errors**. *nature*, **323**(6088):533, 1986. 23

[65] NITESH V CHAWLA, KEVIN W BOWYER, LAWRENCE O HALL, AND W PHILIP KEGELMEYER. **SMOTE: synthetic minority over-sampling technique**. *Journal of artificial intelligence research*, **16**:321–357, 2002. 24

[66] GARY M WEISS, KATE MCCARTHY, AND BIBI ZABAR. **Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?** *DMIN*, **7**:35–41, 2007. 24

[67] CX LING AND VS SHENG. **Cost-Sensitive Learning and the Class Imbalance Problem. 2011**. *Encyclopedia of Machine Learning: Springer*. 24

[68] CHAO CHEN, ANDY LIAW, AND LEO BREIMAN. **Using random forest to learn imbalanced data**. *University of California, Berkeley*, **110**:1–12, 2004. 24

[69] STACEY J WINHAM, ROBERT R FREIMUTH, AND JOANNA M BIERNACKA. **A weighted random forests approach to improve predictive performance**. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **6**(6):496–505, 2013. 24

[70] KRISTIN E BERGETHON, CHRISTINE JU, ADAM D DEVORE, N CHANTELLE HARDY, GREGG C FONAROW, CLYDE W YANCY, PAUL A HEIDENREICH, DEEPAK L BHATT, ERIC D PETERSON, AND ADRIAN F HERNANDEZ. **Trends in 30-day readmission rates for patients hospitalized with heart failure: findings from the Get With the Guidelines-Heart Failure Registry**. *Circulation: Heart Failure*, **9**(6):e002594, 2016. 26

[71] HOLLY C FELIX, BEVERLY SEABERG, ZORAN BURSAC, JEFF THOSTENSON, AND M KATHRYN STEWART. **Why do patients keep coming back? Results of a readmitted patient survey**. *Social work in health care*, **54**(1):1–15, 2015. 27

[72] COLLEEN K MCILVENNAN, ZUBIN J EAPEN, AND LARRY A ALLEN. **Hospital readmissions reduction program**. *Circulation*, **131**(20):1796–1803, 2015. 27

[73] HUAQIONG ZHOU, PHILLIP R DELLA, PAMELA ROBERTS, LOUISE GOH, AND SATVINDER S DHALIWAL. **Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review**. *BMJ open*, **6**(6):e011060, 2016. 27

[74] SENJUTI BASU ROY, ANKUR TEREDESAI, KIYANA ZOLFAGHAR, RUI LIU, DAVID HAZEL, STACEY NEWMAN, AND ALBERT MARINEZ. **Dynamic hierarchical classification for patient risk-of-readmission**. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1691–1700. ACM, 2015. 27, 28

[75] CAO XIAO, TENGFEI MA, ADJI B DIENG, DAVID M BLEI, AND FEI WANG. **Readmission prediction via deep contextual embedding of clinical concepts**. *PloS one*, **13**(4):e0195024, 2018. 27, 28

[76] WAEL FARHAN, ZHIMU WANG, YINGXIANG HUANG, SHUANG WANG, FEI WANG, AND XIAOQIAN JIANG. **A predictive model for medical events based on contextual embedding of temporal sequences**. *JMIR medical informatics*, **4**(4), 2016. 27, 33

[77] YASHAR MAALI, OSCAR PEREZ-CONCHA, ENRICO COIERA, DAVID ROFFE, RICHARD O DAY, AND BLANCA GALLEGO. **Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital**. *BMC medical informatics and decision making*, **18**(1):1, 2018. 27

[78] RICH CARUANA, YIN LOU, JOHANNES GEHRKE, PAUL KOCH, MARC STURM, AND NOEMIE ELHADAD. **Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission**. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015. 28

[79] MEHDI JAMEI, ALEKSANDR NISNEVICH, EVERETT WETCHLER, SYLVIA SUDAT, AND ERIC LIU. **Predicting all-cause risk of 30-day hospital readmission using artificial neural networks**. *PloS one*, **12**(7):e0181173, 2017. 27

[80] SARA BERSCHE GOLAS, TAKUMA SHIBAHARA, STEPHEN AGBOOLA, HIROKO OTAKI, JUMPEI SATO, TATSUYA NAKAE, TORU HISAMITSU, GO KOJIMA, JENNIFER FELSTED, SUJAY KAKARMATH, ET AL. **A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data**. *BMC medical informatics and decision making*, **18**(1):44, 2018.

[81] ALVIN RAJKOMAR, EYAL OREN, KAI CHEN, ANDREW M DAI, NISSAN HAJAJ, MICHAELA HARDT, PETER J LIU, XIAOBING LIU, JAKE MARCUS, MIMI SUN, ET AL. **Scalable and accurate deep learning with electronic health records**. *npj Digital Medicine*, **1**(1):18, 2018. 28

[82] JING ZHAO. *Learning Predictive Models from Electronic Health Records*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University, 2017. 28

[83] RICCARDO MIOTTO, LI LI, BRIAN A KIDD, AND JOEL T DUDLEY. **Deep patient: an unsupervised representation to predict the future of patients from the electronic health records**. *Scientific reports*, **6**:26094, 2016. 33

[84] Y TANG, J CHOI, D KIM, L TUDTUD-HANS, J LI, A MICHEL, H BAEK, A HURLOW, C WANG, AND HB NGUYEN. **Clinical predictors of adverse outcome in severe sepsis patients with lactate 2–4 mM admitted to the hospital**. *QJM: An International Journal of Medicine*, **108**(4):279–287, 2014. 33

[85] YING P TABAK, XIAOWU SUN, CARLOS M NUNEZ, VIKAS GUPTA, AND RICHARD S JOHANNES. **Predicting readmission at early hospitalization using electronic clinical data: an early readmission risk score**. *Medical care*, **55**(3):267, 2017. 33

[86] GLENN SHAFER AND VLADIMIR VOVK. **A tutorial on conformal prediction**. *Journal of Machine Learning Research*, **9**(Mar):371–421, 2008. 33

[87] ZIAD OBERMEYER AND EZEKIEL J EMANUEL. **Predicting the future—big data, machine learning, and clinical medicine**. *The New England journal of medicine*, **375**(13):1216, 2016. 34

[88] RADIM ŘEHŮŘEK AND PETR SOJKA. **Software Framework for Topic Modelling with Large Corpora**. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`. 37

[89] FRANÇOIS CHOLLET ET AL. **Keras**, 2015. 39