Predicting healthcare demand using machine learning on patient data

Lamers, I.C.D.

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Industrial Engineering & Innovation Sciences

Innovation, Technology, Entrepreneurship & Marketing

Master Thesis at VieCuri Medisch Centrum

PREDICTING HEALTHCARE DEMAND USING MACHINE LEARNING ON PATIENT DATA

TUE SUPERVISORS

dr.ir. V. (Vikrant) Sihag
dr. V. (Vahideh) Reshadat
dr. A. (Annelies) Bobelyn

VIECURI SUPERVISOR

G. (Gertjan) Vos MSc

AUTHOR

ing. I.C.D. (Ivo) Lamers

Venlo
23-03-2021

SUMMARY

Hospitals in the Netherlands struggle to provide cost-efficient high quality healthcare to an increasingly aging population. National regulations and profit driven healthcare insurers require hospitals to strike a balance between providing care for everyone while keeping expenses within the set budget. The yearly national healthcare cycle requires hospitals to forecast healthcare demand one year in advance, where deviations in fault margins lead to lower treatment compensation to cover for hospital expenses. The more accurate a hospital is able to forecast next year's healthcare demand, the more treatment compensation is received.

Instead of looking at general historical healthcare consumption, this research approaches healthcare demand on an individual patient level basis by incorporating sequential electronic healthcare records (EHR). The machine learning subfield of deep learning offers the capabilities of automatically detecting sequential healthcare demand patterns unrecognisable by humans and leveraging these uncovered patterns in order to predict future outcomes. Therefore, this research aims to validate the predictive performance of the Doctor AI (Choi et al., 2016c) recurrent neural network (RNN) architecture to leverage sequential temporal relationships stored in patient centred EHRs as to more accurately forecast future healthcare demand. To validate the results of the Doctor AI method, it is applied to seven different models with four different hierarchical levels on a local Dutch hospital's dataset with around 1.8 million EHRs from 240k unique patients.

The results of this research show that only the highest hierarchical level of hospital specialism yields a performance that is good enough to be deployed in forecasting praxis, of 75,4% accuracy@3 and 72,6% recall@3. The lower and most detailed hierarchical levels of healthcare products, ICD10-parents, and ICD10-blocks do not yield high enough predictive performance. The hospital specialism model is able to predict at which of the 21 different specialisms the patient will be treated next, but loses out on the granular information of the lower hierarchical levels. Due to this loss of information, the model is not able to improve the healthcare demand forecasting fault margins. By incorporating the hospital specialism results into the forecasting of the healthcare demand, the model is able to achieve a raw estimate within the right order of magnitude, but deviates more than 10% from the current hospital's praxis on all three analysed cycles of 2017 until 2019.

It can be concluded that machine learning in general has huge potential in predicting healthcare demand and is able to handle the challenges of patient centric EHR data. However, the results of this research show that not every algorithm achieves the same level of performance when applied to a different dataset in a different setting. More research on and comparison of different methods and datatypes is necessary to improve performance on the lower hierarchical levels.

ACKNOWLEDGEMENTS

This report presents the results of multiple months of work and research at VieCuri Medisch Centrum and marks the end of the Master Thesis Project for the Master Innovation Management at Eindhoven University of Technology (TU/e) at the faculty Industrial Engineering & Innovation Sciences (IE&IS) within the department of Innovation, Technology, Entrepreneurship & Marketing (ITEM) in cooperation with the department of Information Systems (IS).

The year 2020 has been an unusual one in many aspects. The COVID-19 pandemic, which is influencing our daily life and habits, has made the Master Thesis Project an even bigger challenge. The process of at the dinner table for several hours at a time, walking around VieCuri wearing a facemask on and the total disruption of social and sport events have had an impact on me personally, but have also contributed to my personal development. Without the help and support from my family and surroundings I would not have been able to come this far. Therefore, I would like to use this opportunity to thank them.

I would like to start by thanking my company supervisor Gertjan Vos and his department at VieCuri for hosting my project, taking the time to answer all questions, and giving me full freedom to use and share their recourses. The social interaction and inclusion in the AFAS ERP project have been of great value to me in a time where social distancing has become the norm. I would also like to thank my TU/e supervisor Vikrant Sihag for taking on this Master Thesis Project with me after weeks of bureaucratic endeavour and Vahideh Reshadat for helping me grasp the basics of a research field I had no experience in. Altogether they have helped me progress from start to finish through many sessions of feedback and discussion.

Next, I would like to thank my parents and Nadja for their tireless encouragements and positive reminders at home through the various ups and downs. Without their social and financial support there would have been no Master study. I would also like to thank Juliën Meijboom, Annelieke Brouwer, and Nicky de Rooij for our weekly online coffee meetings. Going through the same process together created a mutual understanding of the challenges faced and gave some possibilities to blow off some steam.

Finally, I would like to thank Lars Bergh, Surbhi Gupta, Hira Arshat and Yassin Jakani for contributing to the translation of theory to coding and being able to execute a project on datasets of this magnitude. Without them I would not have been able to resolve the coding and computational issues successfully.

TABLE OF CONTENT

LIST OF FIGURES AND TABLES

FIGURES **4**

TABLES **3**

1. INTRODUCTION

The healthcare system in the Netherlands is based on managed competition where everybody is insured under the same minimum basic conditions (Van de Ven & Schut, 2009). Health insurers are obliged to accept all individuals, which ensures that the quality of healthcare is independent of age, income or any other (socio-)demographic aspect (van den Berg et al., 2011). Although the system has shown improvements on efficiency and overall cost over the past years, the Netherlands has one of the highest health expenditures per capita in Europe (Kroneman et al., 2016). Due to its aim of equal care for all, like other OECD (Organisation for Economic Co-operation and Development) countries, the Netherlands faces the challenge of "providing high quality health and long-term care services to an ageing population in a cost-efficient manner" (Schut et al., 2013, p. 2). The biggest challenge Dutch hospitals face is to slow the growth of healthcare expenses to the nationally agreed upon target of 0% by the year 2022 (FMS, 2018) while the cost of healthcare is expected to double over the next decade (Schut et al., 2013), and insurers are leaning towards efficiency and cutting operating costs (Van de Ven & Schut, 2009).

The financial target of 0% growth of expenses is made more difficult as hospitals in the Netherlands are required to budget and therefore forecast healthcare demand one year in advance. Healthcare demand for a hospital is defined as the number of specific treatments (often categorized per specialism) that will be consumed. This means that hospitals need to know one year in advance how many treatments (per specialism) will be sold to the market. As the market is a form of managed competition, the only buyers are the numerous healthcare insurance companies. Based on the forecasted healthcare demand and national normed prices, the hospitals and insurers agree on the maximum number of treatments that will be financially covered for the next year. In agreeing on a set budget one year in advance, the hospital locks in the maximum revenue for that year as the insurance companies will not pay out more to cover expenses. Every patient treatment above the forecasted amount will therefore be treated on the full expense of the hospital. Every patient treatment below the forecasted amount will not be paid out in full by the insurers, again restraining retrieved treatment compensation. Because of this, Dutch hospitals are losing out on potential insurance compensation every time the forecast of healthcare demand is off, thereby directly impacting its operational margins and increasing the healthcare expenses it has set to decrease.

Forecasting healthcare demand accurately requires correct predictions on both the number and type of patients who enter the hospital (Finarelli & Johnson, 2004; Kaplan & Porter, 2011). As healthcare demand is derived from patients who require treatment, forecasting healthcare demand could be developed on an almost patient level basis. There lies a key opportunity in predicting on a patient level, as hospitals are rapidly adopting Electronic Health Records (EHR) systems and the structuring and quantity of patient data are becoming increasingly available (Bates et al., 2014; Wiens & Shenoy, 2018). These EHR systems house a large and complex variety of data, which warrant the use of machine learning applications (Wiens & Shenoy, 2018). The recent increased popularity and adoption of EHR systems and the huge size of daily generated patient data (Bates et al., 2014) can explain the increased interest in machine learning within the healthcare sector. The number of studies where machine learning shows good performance in predicting healthcare demand is growing (Chen et al., 2017; Dahlem et al., 2015; Jensen et al., 2012; Lin et al., 2017; Milovic & Milovic, 2012). This growing interest can

also be recognised in the forecasting of healthcare demand and predict treatment outcomes by machine learning in specific fields like diabetes (Prasad & Agarwal, 2014), heart attacks (Srinivas et al., 2010) and oncology (Chen et al., 2020; IBM Watson for Oncology, 2020 Microsoft Project InnerEye, 2020). But most importantly machine learning is validated in outperforming traditional healthcare demand forecasting methods in areas like blood transfusion (Khaldi et al., 2017), trauma patients (Galatzer-Levy et al., 2014), emergency patients (Zlotnik et al., 2015) and patient key resource demand (Jiang et al., 2017). These applications of machine learning on EHR data show that machine learning algorithms cannot only identify treatment demand, but can also help in identifying previously hidden patterns in the development of and relationship between different diseases. In detecting disease development and the sequential relation between diseases, improvements are made in the predictability and demand of future healthcare, thereby potentially decreasing costs (Bhardwaj et al., 2017; Roysden & Wright, 2015; Srinivas et al., 2010). The potential decrease in cost as has been acknowledged by Callahan & Shah (2017) and Bates et al. (2014) lies in the early identification of potential high-cost and high-risk patients, as they are responsible for the biggest part of the yearly healthcare consumption and cost. Leveraging the knowledge of disease patterns and identifying (potential) high healthcare consumers is becoming increasingly important as this increases the predictability of the yearly healthcare demand. Not predicting future healthcare demand accurately carries the risk for healthcare providers of treating patients at their own cost. Although profit margins are not and should not be the main healthcare incentives, "it is vitally important for healthcare organizations to acquire the ability to leverage machine learning tools effectively or else risk losing potentially millions of dollars in revenue and profits" (Raghupathi & Raghupathi, 2014, p. 2).

Deploying machine learning on EHR data could help to more accurately forecast healthcare demand and decrease healthcare expenditures, both "advancing the field toward precise preventive care to lower overall health care costs and deliver care more efficiently" (Yang et al., 2017, p. 1; Yang et al., 2018, p. 1). Additional better forecasting can also enable hospitals to better balance and better fit their health services to the demand (Soyiri & Reidpath, 2013). Although healthcare data is a large source of future opportunities, healthcare practises are different from profit driven industries. They present researchers with "unique challenges that complicate the use of common methodologies" (Ghassemi et al., 2018, p. 1). To be more concrete, this involves the handling of sensitive patient data, missing data, and the impact of wrong conclusions. However, machine learning shows promising results in better recognising hidden patterns that forecast healthcare demand. In the progress of being integrated into healthcare practises, more research and identification of practical applications is required (Wiens & Shenoy, 2018), especially as modern machine learning techniques on EHR data "have not been widely and reliably used in clinical decision support systems or workflows" (Miotto et al., 2016, p. 1). Therefore, the aim of this research is to test machine learning predictive performance in a local hospital context, in order to contribute to the validation of real-world machine learning application and improve on forecasting performance.

1.1 BUSINESS PROBLEM IN CONTEXT

In order to understand the introduced problem in business context, this research utilizes the data and processes at VieCuri Medisch Centrum in Venlo. VieCuri operates in a region which is predicted to have one of the highest and fastest aging populations in the Netherlands during the next twenty years (De Jong & Van Duin, 2010). More background information about VieCuri Medisch Centrum can be found in Appendix A. Some core numbers on the service area and number of patients over the last three years are displayed in Table 1.

VieCuri Medisch Centrum, as any other Dutch hospital, must budget and forecast the yearly number of patients and their presumed treatments one year in advance to make accurate agreements with the different insurance companies. This means that there is a set number of patients which can consume a set number of treatments each year. Although within a specialism the budget can be reallocated under very strict conditions during the year, the budget is usually fixed for that specialism. If the forecast within a specialism is met exactly, the hospital is paid the full agreed upon yearly amount by the insurance companies. However, if there are less patients treated in a particular year, the hospital is only partially compensated, while on the other hand every treatment over budget is at the full expense of the hospital. This yearly continual balance between budgets that are too low or too high results in the situation where every deviation in fault margin on the forecasted healthcare demand cuts into the profitability of the hospital.

Table 1: VieCuri Medisch Centrum in numbers

Year	Service Population	Patient Consults	Unique Outpatients	Patient Admissions
2017	280.000	333.174	110.566	18.324
2018	280.000	339.926	109.869	18.736
2019	257.190	337.637	109.295	17.796

The first analysis of the problem at hand is the healthcare product forecasting performance over the last three financial years. Three years of data were used as these were available for the research. Besides calculating fault margins per product and specialism, this analysis included the translation of performed treatments towards the missed revenue as a result of that. VieCuri offers over 8.000 different healthcare products ranging from the treatment of a bruised ankle to cornea surgery. Of these 8.000 products, 3.395 have been sold at least once to one of the insurance companies such as VGZ, CZ, or Menzis. Of those 3.395 products, 1.027 accounted for 80% of the revenue.

Table 2: Forecasting performance on the top ten healthcare products

Healthcare Product	Hospital Specialism	2017		2018		2019	
		Number	Budget	Number	Budget	Number	Budget
140301007	Internal Medicine	101	2,2%	-185	-4,1%	-368	-8,1%
131999052	Orthopaedics	47	10,7%	26	5,9%	15	3,4%
131999104	Orthopaedics	-8	-2,0%	44	11,1%	1	0,3%
070401008	Ophthalmology	-172	-3,5%	368	7,5%	91	1,8%
079799020	Ophthalmology	373	2,8%	134	1,0%	993	7,4%
979001219	Cardiology	29	10,0%	3	1,0%	41	14,1%
099899050	Cardiology	31	13,7%	1	0,4%	11	4,8%
099699100	Surgery	29	19,1%	25	16,5%	-24	-15,8%
109999068	Pulmonology	-12	-5,6%	5	2,1%	0	0,0%
028899033	Gastroenterology	-36	-1,8%	1156	57,9%	1037	51,9%

Table 2 illustrates the forecasting performance over the last three years for the top ten most sold healthcare products. The first product (140301007) is performed at the internal medicine hospital specialism and is a treatment for a non-clinical chronic renal insufficiency of the lowest severity, also known as a low severity dialysis. Dialysis treatment for 2017 was forecasted to be performed a total of 4.347 times, but was actually performed a total of 4.448 times, resulting in 101 more treatments during that year or roughly 2 more per week, therewith costing 2,2% more on treatment expenses that year. On the other hand, the same product was performed 185 (3,5 times per week) and 368 (7 times per week) times less than budgeted the two years after, resulting in less compensation on treatment expenses by 4,1% and 8,1% accordingly. The forecasting process on moving averages is executed only once a year for the following one as the healthcare system in the Netherlands functions on a yearly budgeting cycle for all institutes, healthcare providers, and insurance providers. The impact of a doctor who is on long term leave, or spikes of demand in a specific specialism during the year, can therefore only be corrected during the forecast of the next cycle. Due to the complexity of healthcare trajectories a previous year does not always accurately predict the next, as can be seen in the high dispersion of forecasting fault margins in both Table 2 and Figure 1. Figure 1 shows the distribution of fault margins from the aforementioned top 1.027 products (80% of revenue) over the years 2017 to 2019. The fault margin in the individual healthcare products range between -100% until an extreme 41.600% off.

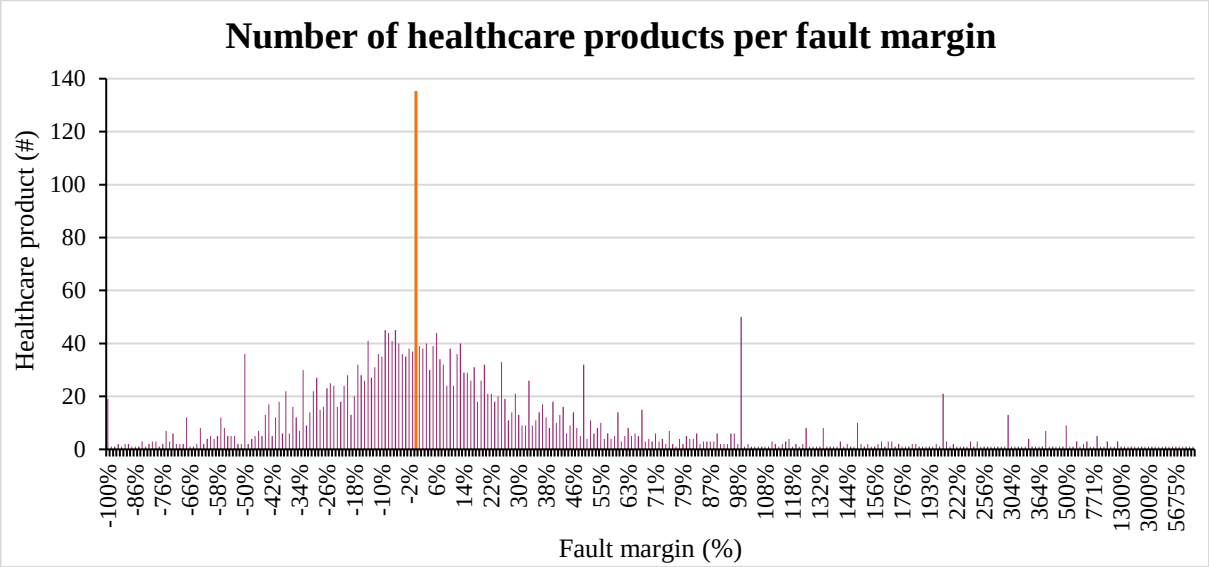

Figure 1: Forecasting performance on the top 1027 healthcare products

Within a specialism there is flexibility in combining specific healthcare product budgets. This means that the overperformance of one healthcare product can be compensated with the underperformance of another, as long as the treatments are offered by the same specialism. For example, if arthrosis treatment of the hips or pelvis is performed more in a single year, but the demand for arthrosis treatment of the knee is low, the budgets can be levelled, as they are treatments which belong to the orthopaedic department and specialism. However, these budget deviations cannot compensate for the under- or overperformance of a bronchopneumonia treatment, which belongs to the pulmonology department and specialism. Therefore, instead of only looking at the error margin of individual healthcare products, it is of higher interest to look

at the aggregated margin of error per specialism. In Table 3 the aggregated too low and too high forecasted fault margins per specialism are shown. The hospital specialisms are ordered on the size of their revenue from highest to lowest. By not zooming in on only the average but at both the too low and too high aggregated forecasting fault margin, this overview gives a better insight into the severity of the forecasting error as it shows treatments which were not compensated (+) and treatments which were not covered in full (-). The average of the fault margin per specialism cannot be recovered and is therefore a direct loss of treatment compensation. For the last three years this has resulted in a loss of €5.129.930 in 2017, €8.065.545 in 2018 and €8.855.869 in 2019.

VieCuri has recognised a problem in the increasing loss of treatment compensation in combination with an aging region and the accompanying expected increase in healthcare costs; on the other hand, the national agreement has pledged to reduce the growth of healthcare expenses to 0% in the year 2022 (FMS, 2018). This combination of factors has contributed to the belief that predicting future healthcare demand in its current yearly moving average format is not sufficient anymore. Being able to predict healthcare demand more accurately is not only the basis for adequate budgeting, but also the foundation of capacity planning and availability of resources (Kaplan & Porter, 2011). As the amount of EHR data, which is stored on a daily basis, cannot be analysed efficiently by hand, VieCuri is interested in learning how to incorporate modern machine learning tools in order to interpret the large amounts of data in a timely and efficient manner in order to better forecast healthcare demand.

Table 3: Forecasting performance per specialism

Hospital specialism		**2017**		**2018**		**2019**	
		Over	**Under**	**Over**	**Under**	**Over**	**Under**
Cardiology	CAR	10,0%	-10,2%	7,7%	-9,4%	11,5%	-7,9%
Surgery	HLK	10,2%	-9,3%	12,1%	-9,3%	9,1%	-12,7%
Internal Medicine	INT	9,0%	-7,3%	11,7%	-9,1%	8,8%	-8,3%
Orthopaedics	ORT	13,4%	-4,7%	12,4%	-10,3%	10,2%	-9,4%
Pulmonology	LON	8,4%	-6,4%	6,1%	-6,1%	7,9%	-6,5%
Ophthalmology	OOG	3,5%	-2,0%	5,5%	-0,6%	8,8%	-1,9%
Gastroenterology	GAS	11,0%	-8,7%	22,8%	-11,1%	22,6%	-10,3%
Gynaecology and Obstetrics	GYN	12,4%	-5,2%	7,4%	-6,0%	8,2%	-10,8%
Urology	URO	9,6%	-4,7%	11,5%	-4,2%	16,6%	-4,0%
Neurology	NEU	8,2%	-6,3%	14,0%	-4,9%	16,8%	-16,1%
Throat, Nose and Ear	KNO	7,6%	-5,5%	8,2%	-6,3%	11,4%	-7,1%
Paediatrics	KIN	21,7%	-14,2%	15,1%	-23,0%	12,2%	-19,9%
Dermatology	DER	4,0%	-6,4%	7,8%	-4,0%	5,7%	-7,8%
Plastic Surgery	PCH	11,8%	-4,5%	15,4%	-5,1%	19,3%	-3,7%
Geriatrics	GER	23,8%	-1,7%	20,0%	-15,8%	0,3%	-14,5%
Rheumatology	REU	7,8%	-5,9%	15,4%	-2,6%	3,2%	-4,2%
Anaesthesiology	ANA	10,3%	-16,7%	29,2%	-10,7%	37,2%	-18,0%
Neurosurgery	NCH	13,6%	0,0%	0,0%	-5,9%	0,0%	-12,2%
Rehabilitation	REV	0,0%	-0,4%	0,1%	0,0%	0,1%	0,0%
Loss of treatment compensation €		**€ 5.129.930**		**€ 8.065.545**		**€ 8.855.869**	
Average fault margin %		**3,0%**		**4,8%**		**5,2%**	

1.2 RESEARCH QUESTIONS

This research aims to answer the question whether machine learning is able to predict healthcare demand more accurately than the VieCuri current practise. Hospitals can use these insights to better predict and manage the budgeting of specific treatments, obtain a more accurate forecast and keep treatment costs affordable. The main research question is therefore as follows:

How can VieCuri Medisch Centrum more accurately forecast healthcare demand applying machine learning on EHR data in order to decrease loss of treatment compensation?

To answer the main research question a set of sub-questions are formulated. First, a literature review is conducted in order to determine which present published machine learning methods would be theoretically most suitable for the application of healthcare product prediction in state progression/time series within the healthcare context. Therefore, the first sub-question is formulated as follows:

Q1: Which machine learning methods would be theoretically most suitable in forecasting healthcare demand according to present published literature?

Secondly, the correct variables needed as input for the selected methods need to be extracted from the VieCuri EHR database. In order to compile the correct dataset from the data warehouse the following sub-question is formulated:

Q2: Which EHR variables need to be extracted from the VieCuri database as input for the identified machine learning methods?

Thirdly, the identified machine learning method is trained and tested on the compiled dataset of variables in order to determine the performance on predicting healthcare demand. To gain knowledge in the performance of the selected method the sub-question is as follows:

Q3: What is the performance of the identified machine learning methods on the available VieCuri EHR dataset?

Finally, the financial impact of incorporating the machine learning method in the forecasting of healthcare demand at VieCuri is discussed. Hence, the final sub-question:

Q4: What would be the financial impact of incorporating the selected machine learning method into the yearly process of forecasting healthcare demand at VieCuri?

1.3 THESIS OUTLINE

Chapter 1 as presented above introduces the subject and business problem of this research and defines the questions which are answered in the following chapters. Chapter 2 focusses on existing literature on the topic of machine learning in healthcare, the nature of EHR data, and the application of deep learning within this field. It also presents the approach that is followed in performing the systematic literature review and thereby answers the first sub-question on most suitable existing machine learning algorithms. In order to generalize the results of the in chapter 2 identified methods in predicting future healthcare demand, this research aims to validate the methods on the EHR data of the local VieCuri Medisch Centrum. The gathered VieCuri data and methodology of the selected machine learning algorithm of this research is therefore discussed in more detail in chapter 3. Chapter 3 thus answers the second sub-question of this research and elaborates on the applied method. The performance metrics, which are applied to evaluate the methods performance, are also introduced in this chapter. In chapter 4 the results of the different machine learning algorithm configurations are analysed and explained; furthermore, this chapter goes more in depth on the potential financial impact of incorporating machine learning in VieCuri forecasting praxis. Therewith, the chapter answers both the third and fourth sub-question of this research. Chapter 5 outlines the conclusion and discussion of this research, which answers the main research question and discusses the limitations, theoretical and practical relevance.

2. THEORETICAL BACKGROUND

In this chapter the first sub-question of the research is answered. The first paragraph of this chapter outlines the complex nature of EHR data as it poses some unique challenges in contrast to other types of data. The following subparagraph 2.1.1 is dedicated to introducing the subject of machine learning and its scope within the field of healthcare, as it identifies some clear distinctions from other fields of application. This section concludes with paragraph 2.1.2 where the subfield of deep learning is elaborated upon, as its advantages can be applied to the problem of diagnosis and healthcare demand prediction. Section 2.1 on EHR data, machine learning, and deep learning together contributes to understanding of the research field and its defining characteristics; the gained knowledge has contributed in both the focus and framing of the right search query for the literature review that follows. The literature review, of which first the methodology is explained in paragraph 2.2 and secondly the assessment of identified methods is explained in paragraph 2.3, contributes to the identification and understanding of existing machine learning methods on diagnosis prediction. By combining the acquired knowledge and the literature review results this chapter concludes with paragraph 2.4, in which the identification of three existing machine learning algorithms is presented. These three methods are capable of solving the problem at hand of predicting future healthcare demand.

2.1 COMPLEXITY OF EHR DATA

Electronic Health Record (EHR) data is the collection of all historical patient files most often stored on an individual patient level basis in large data warehouses. An EHR file is a digitalized representation of patient-centred, real-time records which contains medical information, treatment history, lab tests and other results of any hospital visit or admission. The sources from which EHR data can be retrieved differ largely and are typically compiled from either insurance claims, pharmacy details, local clinic EHR files or when aggregated, a combination of the three. There is no one standard type of EHR data, therefore a whole array of different datasets for predictive models can be recognised. Most often they include patient demographics and either one or a combination of the following data fields: diagnostic codes, procedure codes, clinical notes, stay period, insurance amount claimed, amount reimbursed, self-reported health status, images etc. (Zhao et al., 2005; Pietz et al., 2004). One of the challenges of working with EHR data is the encoding of all these different types of data into an (often vectorised) dataset for analysis purposes. Due to its temporal nature, an EHR patient file typically consists of multiple inpatient records. As previous medical events can have an impact on future visits, capturing the dependencies of these medical records is crucial in predicting future diagnosis (Ruan et al., 2019). In order to create understanding of the process of capturing medical events and visualizing the process of healthcare predictions using EHR data, Solares et al. (2020) included Figure 2 in their research, which illustrates the concept of leveraging vectorized EHR data to predict future healthcare outcomes.

Known issues in handling EHR data are the temporality nature, high dimensionality, large volume, sparseness, incompleteness, noise, random error, systematic bias, and different data types - categorical (procedure codes, diagnosis codes, drug codes, gender), numeric (BMI), time series (sensor data), dates (admission date, discharge date), natural free text (clinical notes), and scans (images) (Ruan et al., 2019; Sheets et al., 2017).

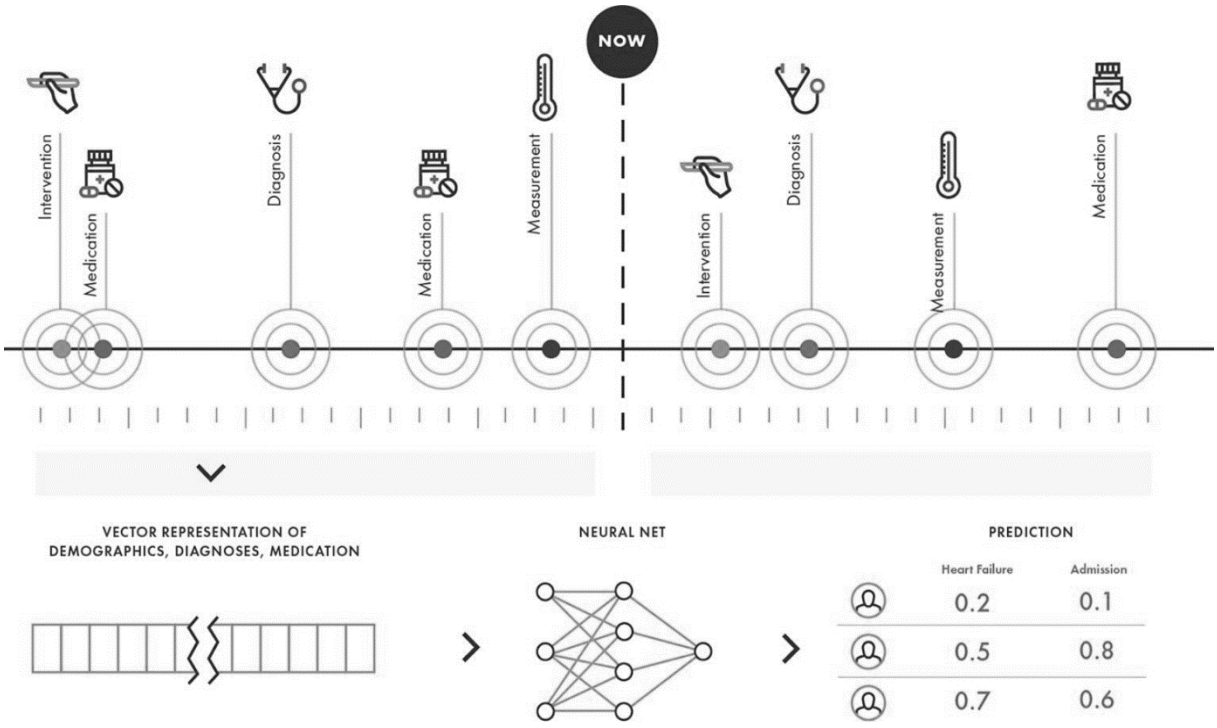

Figure 2: The concept of representation learning to healthcare prediction (Solares et al., 2020 p. 1)

Problems occur when extracted data features with simple encoding schemes are used to represent the heterogeneous nature of EHR data. Hand-crafted schemas like categorization of continuous features, label encoding, frequency, or one-hot-encoding fail to capture the latent similarities between medical concepts. Representation learning methods in contrast create more dense and low-dimensional vectors while capturing the semantics in context. These advantages can be used in machine learning applications. Representation learning is gaining popularity in various fields, as it eliminates the need of time-consuming feature engineering, and it extracts and organizes the similarity and discriminative information from the data in vector forms (Deng et al., 2010; Seidi et al., 2011; Coates & Ng, 2011; Yu et al., 2011; Bengio et al., 2013). Although methods differ from one another (Choi et al., 2016a; Nguyen et al., 2016; Zhou et al., 2017) representation learning can capture the underlying dependencies between visits, diseases, or cluster diagnosis, and thereby overcome the problems of sparse high dimensional medical codes and the struggle of capturing the temporal-hierarchical nature of EHR data. The underlying dependencies in EHR data are especially important in identifying patterns in health development, which are between different healthcare domains, as cause and effect are not always obvious when being limited to a single domain or specialism (Solares et al., 2020).

Due to the complexity and the different policies of different countries, there is not one standardized approach of recording EHR data. Even between hospitals or different hospital specialisations the data that is captured can differ. Therefore, if the data is to be used to draw generic conclusions, acting on the limitations of local data is of importance. Morton et al. (2016) acknowledged the importance of standardized codes for medical data fields and pointed out how several data entries may be deemed useless for data mining applications in absence of these. Although the World Health Organization (WHO) has put great efforts in standardizing medical classification codes for both diseases and procedures (ICD-standard), the implementation of this framework can still vary between countries and institutes. It is therefore

of great importance to understand that applying similar models, even when applied to the same problem, can yield different results between data from different EHR systems (Lin et al., 2017; Morton et al., 2016). As Lin et al. (2017) explained throughout their paper: there is no guarantee a machine learning algorithm performs equally compared to another on any hospital dataset it is applied to. Acknowledged by many of the incorporated researches (Galatzer-Levy et al., 2014; Jiang et al., 2017; Khaldi et al., 2017; Miotto et al., 2016; Prasad & Agarwal, 2014; Roysden & Wright, 2015; Srinivas et al., 2010; Yang et al., 2017; Zlotnik et al., 2015) more validation of methods should be done which only incorporate data from a single data source or single geographic region in order to show generalisable forecasting power on other EHR datasets before qualifying them as good forecasting methods.

As discussed in this paragraph, EHR data can house a wide variety of datatypes, and researchers are faced with challenges in applying their methods due to the complexity of EHR data. Although more validation on present methods is needed to show generalisability of results, complexity of EHR data can be addressed within the domain of machine learning. As introduced in chapter 1 and more elaborated upon in the following paragraphs, machine learning and in particular deep learning is able to identify and leverage previously hidden patterns in EHR data, which contribute in forecasting future healthcare demand.

2.1.1 MACHINE LEARNING IN HEALTHCARE

Machine learning represents a wide field of multidisciplinary applications, which in the basis consists of mathematics, statistics, and computer sciences. The core concept of machine learning is the automated process of learning from data. Instead of explicitly programming rules and making a computer execute those rules, a model is constructed that learns to optimize a mathematical function based on a large amount of data that is fed to a model. Such a data driven approach can be recognised in many different fields of research, such as natural language processing (Yala et al., 2017), image recognition (Rahane et al., 2018), speech to text translation (Chung et al. 2019), and this paper's topic of interest: medical diagnosis. Medical predictions such as diagnosis, disease, and healthcare outcome prediction are the major application scenarios for machine learning methods within in the healthcare domain (Jensen et al., 2012).

In general there are two main approaches to machine learning, which can both be recognised in healthcare predictions. The first and most widely adopted is supervised learning, which involves training algorithms using known examples and labels of medical events such as prediction of mortality rate (Pitocco et al., 2018), readmission rate (Billings et al., 2012; Silverstein et al., 2008; Upadhyay et al., 2019), patient resource demand (Jiang et al., 2017), days of admission (Hachesu et al., 2013; Morton et al., 2014; Xie et al., 2014; Xie et al., 2015), future medical codes (Shi et al., 2018; Shickel et al., 2017), or future costs (Jödicke et al., 2019; Zhao et al., 2005). The second approach is unsupervised learning, where the algorithm explores the data and develops a structure or pattern without knowing the output. Examples of this can be found in discovering risk patients from doctors' notes (Mikolov et al., 2013), lung cancer detection from images (Rahane et al., 2018), or partitioning Alzheimer patients (Alashwal et al., 2019). Although a lot of research (Ahamed & Farid, 2018; Li et al., 2020b; Yoon et al., 2016) has been done on how machine learning can assist doctors in determining treatment or prescriptions for patients by predicting future healthcare demand, potential financial implications lag behind. The potential financial gains of incorporating machine learning in healthcare prediction have

been addressed by some (Bhardwaj et al., 2017; Callahan & Shah 2017; Dahlem et al., 2015; Huff et al., 2018; Panch et al., 2018; Rose 2018 Zeng et al., 2020), however only one of them (Zeng et al., 2020) conveys by how much this impact could potentially be, leaving room for validation of these claims. In order for this research to validate machine learning performance as well as potential financial impact on the current forecasting methodology, the reliance on labelled data is of importance. Therefore, this research relies on the current most dominant machine learning approach of supervised learning.

Leveraging machine learning to accurately predict future patients' healthcare demand would not only assist in future healthcare planning, but also in predicting future expenditure, and in allocating the resources to efficiently optimize cost as a result (Bhardwaj et al., 2017; Yang et al., 2017; Yang et al., 2018). Several studies have concluded that a large percentage of hospital expenditures is contributed to patients suffering from certain chronic diseases, diseases having high procedural cost or resource cost, and patients suffering from multiple diseases, with age also playing a major role (Bakx et al., 2016; Dove et al., 2003; Wammes et al, 2017). Predicting the type of patient and the number of patients falling into these categories are significantly important factors for forecasting healthcare demand and expenditure (Finarelli & Johnson, 2004; Kaplan & Porter, 2011; Xie et al., 2015). In other words, the patients with the highest consumption of care also tend to utilize the highest future consumption of healthcare and are therefore responsible for the biggest part of healthcare expenses.

Traditional methods across the machine learning spectrum (logistic regression, naive bayes, decision trees, random forests, support vector machines (SVM), and multilayer perceptron (MLP)) have been applied within healthcare predictions and more particularly on EHR data. However, these methods have shown that working with EHR data and capturing its underlying dependencies can be challenging (Barati et al., 2011; Kam et al., 2010; Lei, 2017; Morton et al., 2014). As the volume of healthcare data increases on a continual basis ("150 exabytes in the United States alone, growing 48% annually" (Esteva et al., 2019, p. 24)) the more traditional machine learning techniques struggle without expert knowledge in all specific healthcare fields of application. Within the last ten years most of the literature on analysing EHR data has been focused on statistical and machine learning techniques as identified by Murphy (2012). Only in the last few years have deep learning methods overtaken traditional methods as the dominant field of research (Shickel et al., 2017).

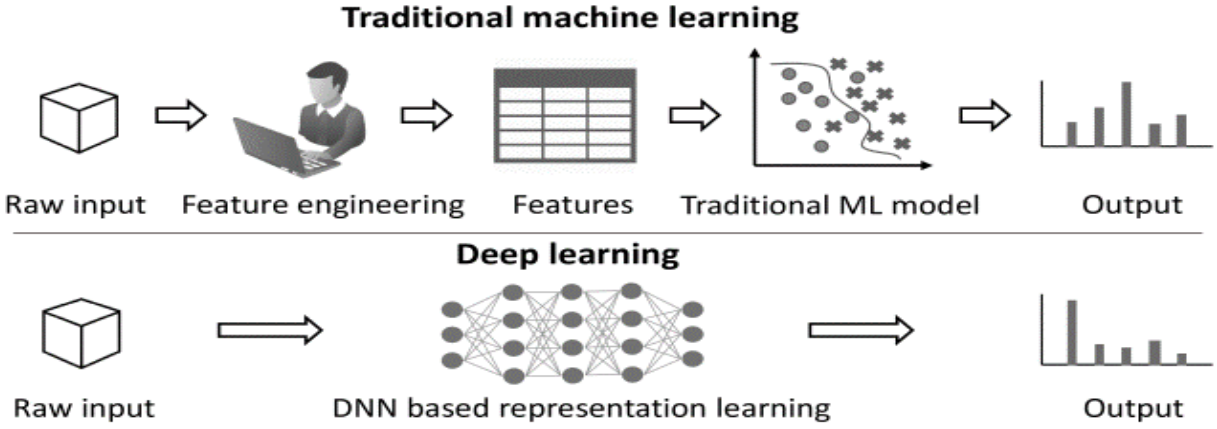

Figure 3: A traditional ML vs DL approach using NN based representation learning (Du et al., 2019 p. 70)

Deep learning, more than the aforementioned traditional methods, can contribute greatly in (1) identifying the underlying dependencies between patient visits and occurring diseases, without the need of expert domain knowledge. Additionally, deep learning utilizes previously hidden patterns (2) to predict which patients will need certain care over an extended period of time. Exactly these patterns over longer time sequences play an important role in predicting future healthcare demand (Solares et al., 2020). Another advantage that deep learning has over traditional machine learning methods is that it does not rely solely on experts feature engineering, but it is able to learn representations in an end-to-end fashion (Du et al., 2019) as is visualized in Figure 3. With traditional methods a doctor or other domain knowledge expert on, for example, cardiovascular diseases is needed to extract and process the features which contribute to the prediction of vein thrombosis in a particular population. Deep learning on the other hand is able to effectively construct the features itself, and relies more on the patterns it discovers in the data. This is particularly an advantage in predicting outcomes across different healthcare domains, as future healthcare demand prediction does.

2.1.2 DEEP LEARNING AND ITS ADVANTAGES

The concept of deep learning is a subfield within the machine learning domain and has seen increased interest over the last few years, driven by the increased capabilities in computational power and the availability of big data. More specifically, in the domain of future healthcare diagnoses, the recent attention can be explained by current achievements in capturing long-range dependencies in healthcare data in an effective manner (Goodfellow et al., 2016; Solares et al., 2020; Xiao et al., 2018). At the core, deep learning is a form of representation learning that is composed of multiple so-called layers, which for the interest of this research consists of layers of patient (data) representations. The deep learning methodologies suitable for diagnosis prediction are most often able to translate EHR patient data in order to develop an often vectorized/binary patient representation, which is then used for pattern recognition. These multiple representation layers are most commonly arranged in sequential order and are composed of a large number of primitive, nonlinear operations, such that the representations of one layer (beginning with the raw patient data) feeds into the next (Esteva et al., 2019). The process where data flows from one layer to the next layer results in an output which is iteratively distorted. This distortion makes the algorithm able to learn highly complex functions as it summarizes how the distances between the embedded points deviate from the original distances between two datapoints.

Traditional methods in healthcare have focused their efforts on analysing the past (after the fact) and rely heavily on expert knowledge of the involved medical variables, as well as on the variable collection and aggregation from data warehouses. Deep learning offers a great advantage in the field of machine learning as it is able to recognize hidden patterns which allow for personalized clinical prediction on a patient level by applying end-to-end models on raw patient data (Gehrmann et al., 2018). Deep learning distinguishes itself here from other methods as it transforms the inputs of an algorithm into outputs leveraging data driven rules which are automatically derived from a large set of patient data, rather than being explicitly defined or able of being detected by humans as is visualised in Figure 3 (Al-Aiad et al., 2018).

Especially within the clinical prediction of the healthcare domain, deep learning methods have shown to be both flexible (by incorporating multiple different types of raw patient data) and capable of handling large and complex datasets (longitudinal event sequences and continuous monitoring data) (Xiao et al., 2018). “It is widely held that 80% of the effort in an analytic model is pre-processing, merging, customizing, and cleaning datasets” (Rajkomar et al., 2018, p. 1). Deep learning requires structural represented data in order to leverage its capability in identifying hidden patterns. This offers an advantage over other predictive modelling techniques as those require the custom creation of a dataset for the specifically intended outcome. Therefore, the scalability of deep learning is not limited when compared to the more traditional methods (Rajkomar et al., 2018; Xiao et al., 2018). Doctors and other domain experts in a certain healthcare specialism are specialised in, and therewith limited to, a particular part of the human body. This creates a well-known gap of knowledge on the different dependencies between diseases from different specialisms, or on the recognition of visiting patterns when not limited to the doctor’s expertise. Deep learning is not concerned with or limited to the domain experts’ expertise on a single disease, specialism, or gland, and is more capable of leveraging the cross domain existing patterns previously hidden (Liang et al., 2019; Shickel et al., 2017; Choi et al. 2016c).

The argued distinctive features deep learning brings to the forecasting of healthcare demand are (1) the capability of automatically detecting patterns unrecognisable by human detection and (2) leveraging these uncovered patterns in order to predict scalable future outcomes. The combination of these two features has shown to drive better performance than traditional machine learning methods within the healthcare domain. Simultaneously, deep learning requires less pre-possessing of patient data and expert domain knowledge for feature engineering, which both can be recognised as time-consuming endeavours considering the size and complexity present in EHR data (Khaldi et al., 2017; Galatzer-Levy et al., 2014; Jiang et al., 2017; Shickel et al., 2017; Zlotnik et al., 2015). When presented with enough computational power to execute the model from end-to-end, deep learning methods can accelerate the process of accurate healthcare demand forecasting and assist in many other medical applications of detecting patterns or predicting probabilities.

2.2 LITERATURE REVIEW METHODOLOGY

The goal of the conducted literature review is the identification of suitable machine learning algorithms capable of predicting healthcare demand across multiple domains and different diagnoses and illnesses. The last three paragraphs on EHR data, machine learning, and deep learning combined provide the context for a more narrowed and focused literature review. In order to find previously suggested algorithms, the literature review uses present systematic reviews on the topic of deep learning on EHR data in order to formulate the relevant search queries. four recent comparative systematic reviews into healthcare demand or diagnosis prediction within the deep learning domain (Al-Aiad et al., 2018; Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018) were the starting point of the executed literature study. The results of the studies differ, as was anticipated, but overlap can be recognised in the array of discussed methods. Leveraging the information of these four studies, the final search query was constructed. The search query was executed in the databases of Web of Science (WoS), Scopus and IEEE/IES Xplore (IEEE) and included only published studies. The inclusion criteria for extracting the articles from the databases are shown in Table 4 and explained in more detail in Appendix B. In Figure 4 a visualisation of the followed approach for the literature review is shown, and a more detailed version is also explained in Appendix B. All the methods which reached the final assessment were used in one iteration of backwards snowballing (Wohlin, 2014) in order to discover any missed methods, which were also assessed on the basis of the inclusion criteria as described in Table 4.

Table 4: Inclusion criteria for the literature review per phase

Phase	**Inclusion Criteria**
Initial Search Query	* Language is (English OR German OR Dutch) * Timespan of online publishment is (∞ ; 11-2020]
First Assessment Reading of Title & Abstract	* Article presents a method that predicts any diagnoses/diseases/healthcare outcomes * Article is available online in full text
Second Assessment Scanning of Abstract, Data Descriptives, Conclusion, Discussion & Implications	* Method has shown performance on predicting multiple diagnoses or diseases simultaneously * Method is validated on a real patient dataset
Final Assessment Reading of Full Article	* Method incorporates the factor time as output in order to restrict diagnoses/diseases in yearly prediction

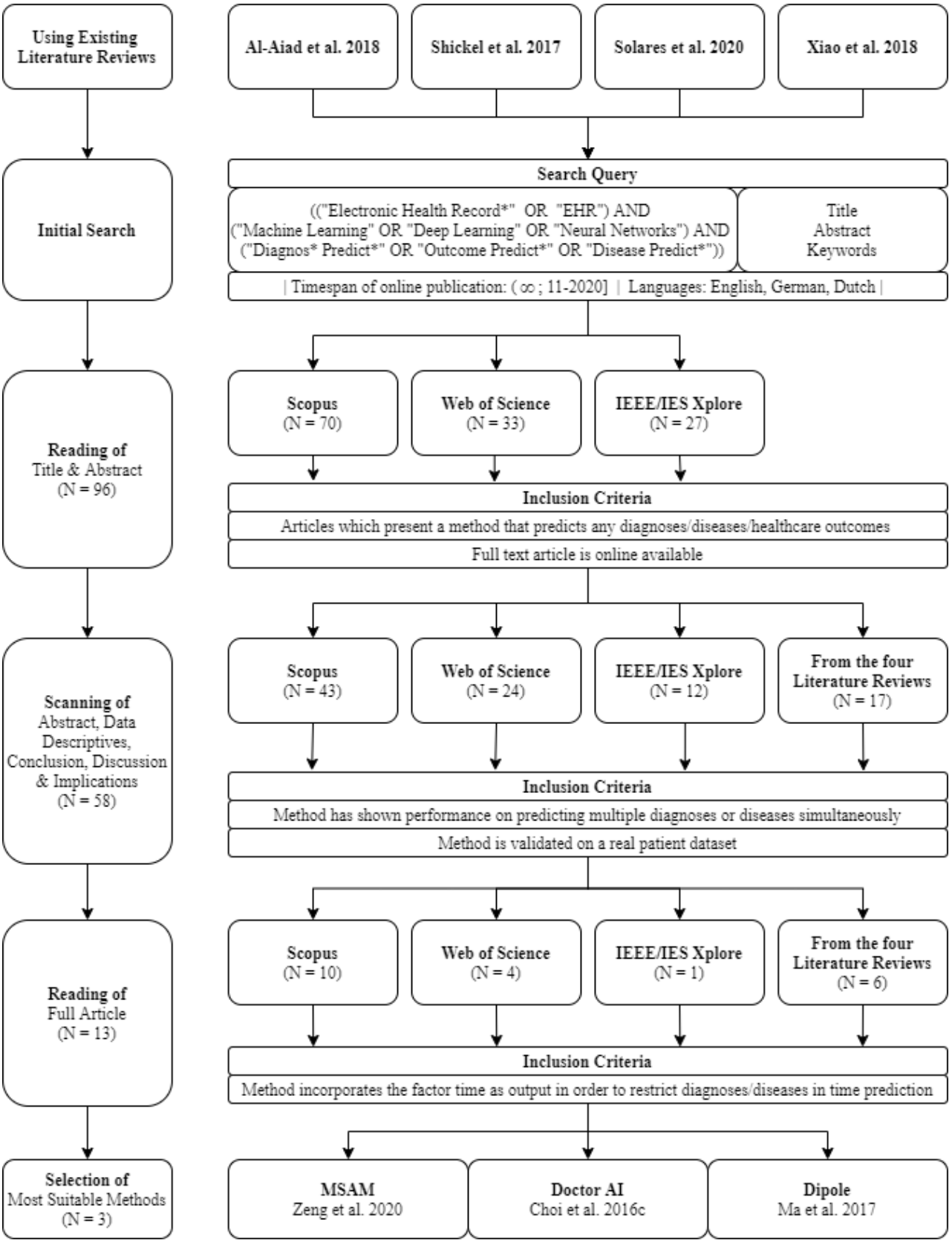

Figure 4: Step by step approach of the literature review

2.3 ASSESSMENT OF IDENTIFIED MACHINE LEARNING METHODS

The results of this literature review are a combination of articles found in the three databases and the articles which were presented in the four aforementioned literature reviews. After the initial search query through the three databases the following number of articles were identified: Scopus 70 articles, Web of Science 33 and IEEE/IES Xplore 27. Due to a certain overlap of the three databases this resulted in a total of 83 unique articles. As Figure 5 illustrates with the distribution of publications within the domain of EHR predictions, most articles were published very recently. It is therefore no surprise that the publication years of these 83 articles are also relatively recent ranging from 2014 until 2021. 13 unique articles, which were presented by the four incorporated literature reviews, were added to this set retrieved from the search query in the next round of assessment. After this first assessment a total of 58 unique articles were considered relevant in predicting some form of disease, diagnosis, or clinical outcome and included the total set of the three databases and four literature reviews. From these 58 unique articles 45 were dropped, mainly for the reason that the presented method was validated on a specific disease or binary outcome prediction. Although some of these methods could be applied to more than one diagnosis cluster, they were not validated on more than one outcome and were therefore disregarded in this research. To give some examples: this included methods for coronary heart disease (Du et al., 2020), sceptic shock (Lin et al., 2018), length of stay & readmission time (Huang et al., 2016), or hospital cost & length of stay (Feng et al., 2017). The 13 remaining articles were analysed in full and were considered for the problem at hand. These methods are summarized in Table 5 listed below. One iteration of the by Wohlin (2014) introduced method of backward snowballing was performed on all the references of the 13 articles in order to identify any other machine learning method. Although both deep learning and traditional methods (such as random forests, SVM, and rule-based) were considered from this first reference iteration, none of them would satisfy the in Table 4 described inclusion criteria. As the backward snowballing delivered no method to be included, the 13 articles listed in Table 5 are considered the final set for this research.

The first main distinction which can be drawn from the identified methods is the type of tasks they are performed on. The majority of articles (six) relied on supervised learning where both the labelled input and output were known. This concerned the methods Doctor AI (Choi et al. 2016c), LSTM with dropout and target replication (named LSTM-DO-TR) (Lipton et al., 2016), Dipole (Ma et al., 2017), RNN (LSTM) with multiplicative attention mechanism (named RNNY) (Mu et al., 2018), RNN (GRU) with demographic information (named RNN-INFO) (Wang et al., 2018), and Multilevel Self-Attention Model (MSAM) (Zeng et al., 2020). The second most dominant type of task which was identified concerned unsupervised learning. In most cases the unsupervised ancestors or neighbouring diseases clustering was the distinctive feature to identify them as such. The following five methods are regarded as unsupervised: Med2Vec (Choi et al., 2016a), Graph-based Attention Model (GRAM) (Choi et al., 2017), Graph Neural networks based Diagnosis Prediction (GNDP) (Li et al., 2020a), Knowledge-based Attention Model (KAME) (Ma et al., 2018), and Deep Patient (Miotto et al., 2016). To conclude, the two most recent published methods Fusion of CNN, BLSTM and Attention Mechanisms (named FCNBLA) (Wang et al., 2020), and Heterogeneous Graph Learning Model (HGM) (Wanyan et al., 2020) performed a semi-supervised task.

Table 5: Machine learning methods which predict future healthcare demand

Author, Year	Name	Method	Approach	Dataset, Population	Pre-Processing	Input	Output	Training and Test	Validation Results	Source*
Wang et al. 2020	FCNBLA	Semi-Supervised; DL	CNN; Bi-LSTM; Attention	Yang Data, 18k, 10 codes	Word embedding of patient representation	Vectorized word embedding of medical records and notes	Predict future diagnosis	Train: 70% Test: 20% Val: 10%	Dx: Recall/macro 0.905 Precision/macro 0.923 F-score/macro 0.913 Accuracy 0.928	Scopus
Wanyan et al. 2020	HGM	Semi-Supervised; DL	FFNN; Graph	MIMIC-III, 4k, 3k codes	Vectorized graph embedding	Vectorized graphical patient representations	Predict future diagnosis	Cross validation method	Dx: F-score 0.751 AUC 0.834	Scopus
Li et al. 2020a	GNDP	Unsupervised; DL	GNN (CNN); Attention; Graph	MIMIC-III, 7k, 5k codes, Dataset-II, 14k, 5k codes	Binary vectorization of representation	Spatial & time-ordered sequence of patient visits & medical code ancestors	Predict future diagnosis	Train: 75% Test: 10% Val: 15%	Dx Accuracy@30 MIMIC-III: 0.863 Dataset-II: 0.924	Scopus
Zeng et al. 2020	MSAM	Supervised; DL	FFNN; Attention	MIMIC-III, 38k, 5k codes; PFK, 146k, 7k codes	Diagnosis and procedure sequences	Vectorized time-ordered sequence of patient visits & Time embedding	Predict future diagnosis, diagnose in next year & cost	Train: 80% Test: 20%	Dx: Recall@30 0.683 DxTx: Recall@30 0.795 $MAE 847.7	Scopus
Ma et al. 2018	KAME	Unsupervised; DL	RNN; GRU; Attention	Medicaid, 99k, 10k codes	DAG; attention + knowledge vectors	Vectorized time-ordered sequence of patient visits & medical code ancestors	Predict future diagnosis	Train: 75% Test: 10% Val: 15%	Dx: Accuracy@30 0.894	Scopus; WoS
Mu et al. 2018	RNNY	Supervised; DL	RNN; LSTM	Haikou People H, 5k, 108 codes	Binary vectorization of representation	Vectorized time-ordered sequence of patient visits	Predict future diagnosis & medication	Train: 70% Test: 20% Val: 10%	Dx: Accuracy@10 0.648	Scopus
Wang et al. 2018	RNN-INFO	Supervised; DL	RNN; GRU	China province, 29k, 1k codes	Vectorisation of diagnosis and medicine	Vectorized time-ordered sequence of patient visits & patient demographics	Predict future diagnosis in 3/6/9/12 months	Train: 80% Test: 20%	Dx: F-score 3month 0.767 6month 0.743 9month 0.816 12month 0.854	Scopus; WoS; IEE
Ma et al. 2017	Dipole	Supervised; DL	BRNN; GRU; Attention	Medicaid, 148k, 1k codes	-	Vectorized time-ordered sequence of patient visits	Predict future diagnosis	Train: 75% Test: 10% Val: 15%	Dx: Accuracy@30 0.836	Scopus; WoS; LR(2)
Choi et al. 2017	GRAM	Unsupervised; DL	RNN; GRU; Graph; Attention	Sutter, 258k, 10k codes	DAG; attention ancestor embedding	Vectorized time-ordered sequence of patient visits	Predict future diagnosis	Train: 75% Test: 10% Val: 15%	Dx: Accuracy@5 (0-20) least common 0.004 (20-40) 0.299 (40-60) 0.422 (60-80) 0.419 (80-100) 0.490	Scopus; WoS; LR(2)
Miotto et al. 2016	Deep Patient	Unsupervised; Ensemble	SDA; RF	Mount Sinai, 705k, 60k descript	None, raw patient data.	Demographics, diagnoses, procedure, lab tests, notes	Predict future disease	Train: 23 years Test: 1 year	Dx: Accuracy 0.929 AUC-ROC: 0.773 F-Score: 0.181	Scopus; LR(3)
Choi et al. 2016c	Doctor AI	Supervised; DL	RNN; GRU	Sutter, 264k, 1k codes; MIMIC-II	Skip-Gram based vectorizing	Vectorized time series of diagnoses and procedure codes	Predict next diagnosis & visit time	Train: 85% Test: 15%	Dx: Recall@30 0.796 DxTx: Recall@30 0.725	LR(3)
Lipton et al. 2016	LSTM-DO-TR	Supervised; DL	RNN; LSTM	LA Children H, ???k, 128 codes	Hand engineered features	Multivariate time series of 13 variables	Predict future multilabel diagnoses	Train: 80% Test: 10% Val: 10%	Dx: AUC Micro 0.856 Macro 0.808 F1 Micro 0.294 Macro 0.149	LR(2)
Choi et al. 2016a	Med2Vec	Unsupervised; NN	FFNN	CHOA, 550k; 10k codes; CMS 831k 14k codes	Binary vectorization of representation	Vectorized time series of diagnoses and procedure codes	Predict future diagnosis	Train: 75% Tests: 25%	Dx: Recall@30 0.757	LR(2)

**From WoS, From Scopus, From IEEE, From the four literature reviews (LR)*

Articles incorporating some variant of Recurrent Neural Networks (RNN) were in the majority with the seven following methods: Doctor AI (Choi et al. 2016c), GRAM (Choi et al., 2017), LSTM-DO-TR (Lipton et al., 2016), Dipole (Ma et al., 2017), KAME (Ma et al., 2018), RNNY (Mu et al., 2018), and RNN-INFO (Wang et al., 2018). The build-up of the methods, however, differs, as some prefer the implementation of so-called Gated Recurrent Unit (GRU) (Choi et al., 2016c; Choi et al. 2017 Ma et al., 2017; Ma et al., 2018; Wang et al., 2018) where others use the Long Short Term Memory Unit (LSTM) variant (Lipton et al., 2016; Mu et al., 2018).

The incorporation of attention mechanisms in representational learning has been increasing over the last few years, so it is therefore not unexpected that six articles used a form of attention mechanisms in their proposed methods: GRAM (Choi et al. 2017), GNDP (Li et al., 2020a), Dipole (Ma et al., 2017), KAME (Ma et al., 2018), FCNBLA (Wang et al., 2020), and MSAM (Zeng et al., 2020). Three articles incorporated some variant of Artificial Neural Network (ANN), a so-called Feedforward Neural Network (FFNN). Their implementation, however, differs completely, as Med2Vec (Choi et al., 2016a) used binary vectorization of patient visits whereas MSAM (Zeng et al., 2020) used attention mechanisms, and HGM (Wanyan et al., 2020) relied on graphical representation learning. The only other method which relied on graph representation learning is GNDP (Li et al., 2020a). A very unique approach is the very recently published FCNBLA (Wang et al., 2020) which leveraged advantage from multiple directions by incorporating the following three in parallel: Convolutional Neural Networks (CNN), Bi-directional LSTM, and attention mechanisms. The Deep Patient method of Miotto et al. (2016), which uses Stacked Denoising Autoencoders (SDA) followed by a Random Forrest approach, is the only traditional method and methodology that did not rely on neural networks. They showed that their representation learning technique performed better than shallow feature learning techniques on predicting future disease, and are for that reason used as a benchmark or starting point in many of the deep learning research papers that followed (Choi et al., 2016c; Choi et al., 2017; Ma et al., 2018; Mu et al., 2018; Wang et al., 2018; Zeng et al., 2020).

The proposed methods utilize either publicly available databases (MIMIC-II or MIMIC-III) in combination with a local hospital or regional dataset, or one or the other. Great differences can be recognised in the volume of patient data which is incorporated and the number of codes which are predicted. This makes the results difficult to compare with one another, as it is clear that predicting 10 different codes from 18k different patients (FCNBLA (Wang et al., 2020)) yields different results than predicting 1.183 codes simultaneously from 264k different patients (Doctor AI (Choi et al., 2016c). Noteworthy to mention is that none of the articles relied on European retrieved datasets; in fact, all datasets are either based on Chinese or American patient data.

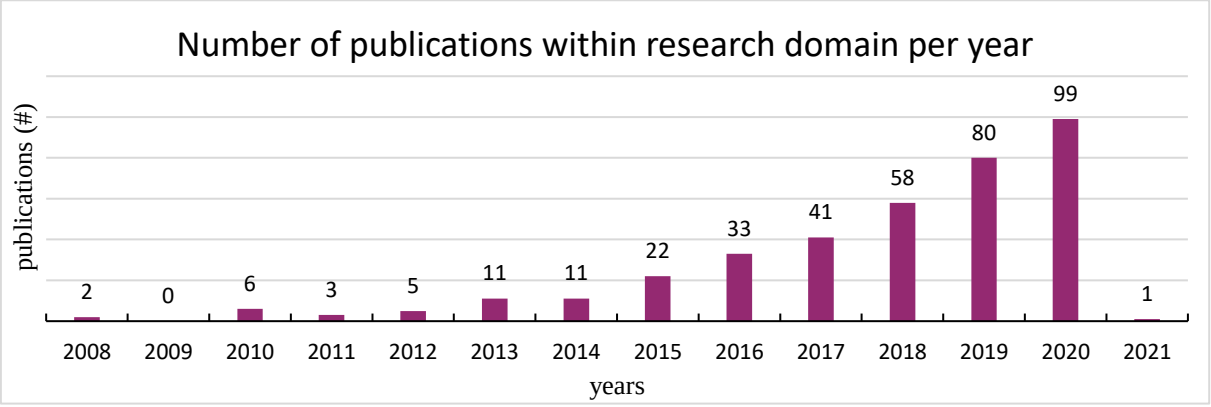

Figure 5: Distribution of domain publications ("EHR" AND "Prediction" AND "Health")(PubMed, 2020)*

2.4 DISCUSSION ON SUITABLE MACHINE LEARNING METHODS

The result of the literature review as provided in Table 5 includes, as assessed in paragraph 2.3, a variety of different machine learning methods, which all incorporate some form of the earlier explained subfield of machine learning. At the core all methodologies, although presented in different shapes and using different mechanisms, leverage patient data in order to predict future medical events. In order to gain a better understanding of the mechanisms, some of the most promising methods based on their capability of solving the problem of forecasting future healthcare demand are discussed below. As this research is concerned with solving a real-world problem, the practical applicability of the method is of high importance.

In general the presented methods rely on some form of chronologically ordered medical information from patients' history in order to create sequenced patient representations. RNNs have been widely adopted to leverage such chronologically ordered information as both forms of LSTM and GRU exhibit temporal dynamic behaviour such as Doctor AI (Choi et al., 2016c), LSTM-DO-TR (Lipton et al., 2016), RNN-INFO (Wang et al., 2018), and Dipole (Ma et al. 2017). However, RNN models which follow such a full black box approach are more difficult to interpret.

To tackle the interpretation issue, Choi et al. (2016a, 2016b, 2017) added weight to the interpretation of deep learning by applying the Skip-Gram embedding method as introduced by Milokov et al. (2013). This method makes it possible to create continuous vector representations of words, or in the medical field most often medical codes, which are scalable without losing interpretability and can be used as input for deep learning models. First, Choi et al. (2016a) built on this by using 2-level representation learning to capture the similarity between different medical codes within a visit, as well as between co-occurrence of these codes, calling it Med2Vec. Although Med2Vec encapsulates the aspects of EHR data which are used to predict future medical events, it doesn't incorporate the dimension of time, or represent the underlying medical relationships. Although they don't fit the problem at hand, Med2Vec, as well as Deep Patient, have proven to be robust baselines for further research (Choi et al., 2017; Ma et al., 2017; Ma et al., 2018; Mu et al., 2018; Yang et al., 2019). Second, Choi et al., (2017) used their graph-based attention model (GRAM) to represent each visit of a patient, overcame the issue of data insufficiency, and included domain knowledge. KAME (Ma et al., 2018) built on the foundation of GRAM's knowledge graph embedding by also including knowledge of ancestor medical codes, but both failed to capture the relationship between diagnoses and healthcare products, and they perform best in niche applications. Knowing what a particular patient's future healthcare demand is, is valuable and relevant to a great spectrum of research done in healthcare and can have different approaches.

The scope of this research is to leverage the historical relationships between visits on an individual patient basis to predict future healthcare demand of the whole service population for the next year. Therefore, in order to make the results of the research usable in practice, the scope of the relevance is restricted to machine learning methods which are capable of including some dimension of the time variable. The exact interpretability of the results, although considered, is therefore less important than it would be on an individual patient level basis.

2.4.1 SELECTION OF THE INCLUDED METHOD

The three methods most closely resembling the problem of predicting future healthcare demand are Doctor AI (Choi et al. 2016c), Dipole (Ma et al., 2017), and MSAM (Zeng et al., 2020). All three incorporate mechanisms to capture the underlying relationships present in time sequenced EHR data on a patient level. A comparison of these three methods is displayed in Table 6.

Table 6: Comparison of Doctor AI, Dipole, and MSAM

	Doctor AI	**Dipole**	**MSAM**
Method	Supervised, DL	Supervised, DL	Supervised, DL
Architecture	RNN, GRU	BRNN, LSTM, Attention	FFNN, Attention
Source of data	Sutter Health, Palo Alto Medical Foundation, California, USA	Medicaid Claim Data, USA	Partner for Kids Paediatric, Ohio, USA
Number of patients	263.706	147.810	146.287
Excluded patients	Less than 2 visits	Less than 5 visits	Less than 2 visits
Timespan	8 years	1 year	2 years
Number of visits	14.400.985	1.055.011	1.301.954
Number of codes	38.594	8.522	12.334
Predicted classes	1.183	426	280
Train, Test, Validation	85%, 15%, 0%	75%, 10%, 15%	80%, 20%, 0%
Doctor AI recall@30	79,58%	-	-
Dipole accuracy@30	-	84,75%	-
MSAM recall@30	78,87%	78,80%	79,48%
Advantages	+ High performance on large multiclass classification problems	+ Higher capability of transferring past information than RNN.	+ Besides code also capturing visit level relationships + Better at handling irregular time intervals
Disadvantages	- Requires high volume per disease. - Suffers performance loss at patients with low number of visits	- Difficulty handling irregular time intervals - Only results on 1 year of data, with exclusion of low visit patients (cherry picking)	- Dependent on precalculated cost weights - Only results on 2 years of data

Although suffering from interpretability issues, RNNs are still the most widely adopted methods for diagnosis and prognosis prediction. Doctor AI, as introduced by Choi et al. (2016c), is a method for predicting future medical events of patients using the GRU architecture of RNN. Doctor AI utilizes sequences of time and event pairs occurring in the patient's timeline across multiple treatments as input. With this method Choi et al. (2016c) improved on the Skip-Gram based medical representation by incorporating the factor of time in order to not only predict in which cluster of diagnoses a next patient visit will fit, but also the time in-between. Doctor AI showed a comprehensive and robust method on a relatively large number of patient data (264k), not being restricted to only a few diseases or diagnosis clusters. Although this is an important feature for the problem at hand, it is also acknowledged to have lesser performance on diseases with low representation in the data.

Ma et al. (2017) proposed, not long after Doctor AI, their method Dipole, that uses an RNN variant of attention based bidirectional LSTM to significantly improve the prediction accuracy compared to the state-of-the-art diagnosis prediction approaches. Dipole has a simpler architecture and models the patient visits in only time-ordered or reversed time-ordered sequence, whereafter it employs attention mechanisms in both directions. The model was trained on the Diabetes and Medicaid dataset. Though the Medicaid dataset consisted of a significantly greater number of patients (148k), the number of diagnoses and procedure categories was limited and similar in both datasets, making the results more difficult to generalize over an entire hospital population. Nevertheless, compared to other baseline methods it yielded slightly better results.

Zeng et al., (2020) very recently solved a problem similar to the one at hand with its proposed method called Multilevel Self-Attention Model (MSAM), which besides predicting future diagnoses also clustered them to predict yearly medical needs, as well as future medical cost. In this research Dipole, Doctor AI, and multiple other baselines were compared on the same dataset, showing slightly better results for MSAM. MSAM, in contrast to the other methods, is more capable of capturing the underlying relationships, temporal information, and dependencies among visits as it has less difficulties handling irregular time intervals between visits. However, by showing these results on a dataset with datapoints over only a two year timespan, the results need to be more validated in order to show whether the attention model they propose is an improvement on the BRNN (Dipole) or GRU (Doctor AI) architecture for predicting future healthcare needs.

The three methods Doctor AI, Dipole and MSAM have overall contributed to improved prediction on EHR data, dealing with aspects such as temporality and high dimensionality in a supervised deep learning task. However, access to hospital data is often restricted to one specific research, making it difficult to validate the findings of these methods on different datasets. In order to generalize the results of the methods in predicting future healthcare demand this research aims to validate the methods on the EHR data of the local VieCuri Medisch Centrum. As the nature of this research is restricted in time and knowledge to model a method from scratch, it has been decided that practical validations can only be executed on one of the methods. In this regard, both Doctor AI and MSAM codes were made publicly available. However, only Doctor AI fully revealed the coding mechanisms publicly and resembles the dominant methodology in this research field. Therefore, Doctor AI is incorporated in the methodology of this research.

3. METHODOLOGY

The methodology and planning for this research followed the Cross Industry Standard Process for Data Mining (CRISP-DM) framework (Shearer, 2000). The CRISP-DM as displayed in Figure 6 serves as an overarching research umbrella of which the following phases are covered in this research: Business understanding (1), Data understanding (2), Data preparation (3), Modelling (4) and Evaluation (5), therewith leaving out Deployment (6). The phases Business understanding and Data understanding were extensively covered in the previous two chapters, defining the scope and context of the research and business problem. In this chapter the third and fourth phase of data preparation and modelling are covered, as well as the evaluation metrics, which will be used in the last phase of evaluation. Finally, in the next chapter, the results of the models are evaluated, thereby concluding the CRISP-DM phases relevant to this research.

The first paragraph of this chapter explains the executed data preparation and cleaning, and it gives an overview of all relevant variables and the executed pre-processing for the final EHR dataset. This paragraph thus answers the second sub-question of this research on the available data variables. The process of data gathering was executed in parallel with the literature study conducted in the previous chapter. The second paragraph is dedicated to the machine learning algorithm Doctor AI (Choi et al., 2016c) which is used as deployment for the CRIPS-DM modelling phase of the research. The paragraph goes more in depth on the architecture, mathematical functions and distinctive features of the method. In the third and final paragraph an explanation is given on the relevant performance metrics which are used in the following chapter to evaluate the method on its performance on the VieCuri dataset.

Figure 6: The CRISP-DM Framework (Smartvision, 2019 p. 1)

3.1 DATA EXTRACTION AND PREPARATION

The source data for this research was retrieved from the VieCuri Medisch Centrum data warehouse and consists of almost 4 million rows of so-called *diagnose-behandel-combinaties* (DBC), which were extracted on a patient level basis. A DBC in the Netherlands is a predefined combination of diagnosis and treatment plan by the Dutch Healthcare Authority. After removing all duplicates, empty registrations, or very obvious outliers (patients with an age of -5 or 136 years old) a total dataset of 1.8 million rows off DBCs on more than 240k patients was compiled. The reduction of almost half of the number of DBCs is mostly due to follow-up DBCs being automatically generated in the system after the first one closes. All empty automatically generated DBCs which remain empty are therefore disregarded for this research. The data spans over eight years from January 2012 to December 2019. The decision to not incorporate data before the year 2012 was made, because this data could not be systematically retrieved in a complete manner. The decision to not incorporate data after 2019 was made due to a drop in the number of patient visits and variety of offered treatments as a result of the COVID-19 pandemic (which as acknowledged by the Dutch Authorities started in the Netherlands on 27th February 2020 (RIVM, 2020)).

The dataset contains patients ranging from the age of 0 until the age of 106 and is, with 49,3% male and 50,7% female, representative of the average Dutch population during this time span (49,6% male and 50,4% female by CBS, 2020). The distribution of the number of DBCs per age and involved cost per age are displayed in Figure 7 and Figure 8. The extracted variables of the dataset are discussed in more detail in paragraph 3.1.1 whereafter the pre-processing of the data is elaborated upon in paragraph 3.1.2.

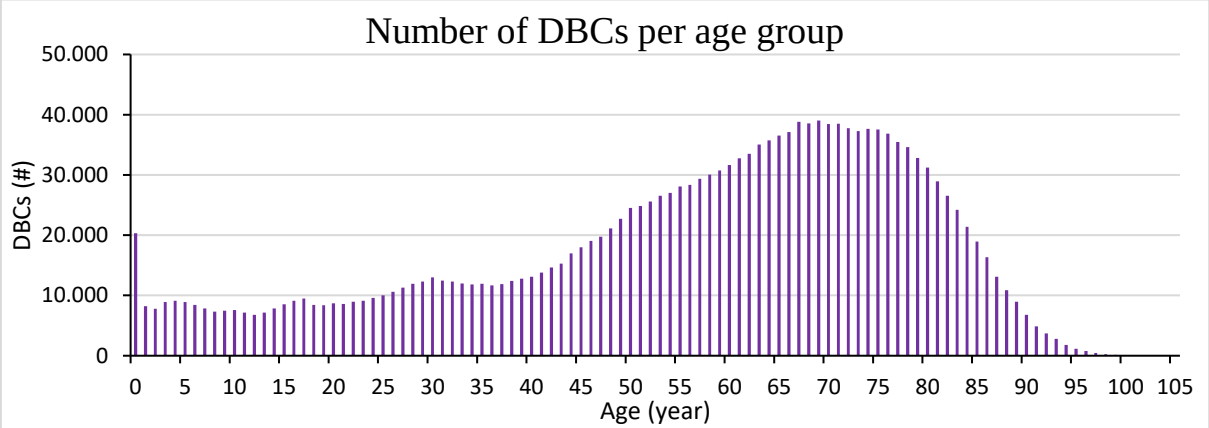

Figure 7: Distribution of number of DBCs per age

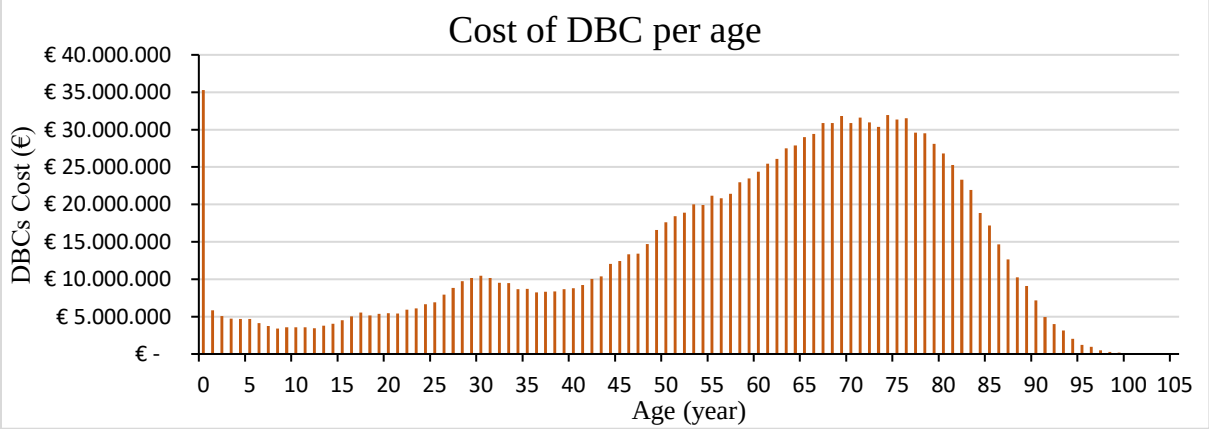

Figure 8: Distribution of cost per age

3.1.1 SELECTED VARIABLES

The extracted dataset includes a total of 11 variables, which through adding and expanding some of the features concludes in a total of 16 variables, summarized in Table 7. The five variables that are added concern the four hierarchical levels of ICD10-codes which are linked to the hospital diagnosis code and the DBC cost, which in turn are linked to the healthcare product and declaration codes. The gathering of data has been executed in parallel with the literature study in chapter 2. Therefore, the dataset also incorporates variables which in the end where not leveraged by the method explained in paragraph 3.2, such as age, gender, declaration code, and cost, but which would have been leveraged by other identified methods. In this paragraph the linkage between the different variables is discussed in more detail.

Table 7: All included variables in the final dataset

Name	Data Type	Unique Values	Labelled Data
DBC_ID	INT (7)	1.797.467	100,0%
EPISODE_ID	INT (7)	1.029.343	100,0%
PATIENT_ID	INT (7)	240.841	100,0%
GENDER	BINARY (M = 1 | V = 0)	2	100,0%
BIRTH_DATE	DATE (DD-MM-YYYY)	35.806	100,0%
STARTDATE_DBC	DATE (DD-MM-YYYY)	2.922	100,0%
ENDDATE_DBC	DATE (DD-MM-YYYY)	3.041	100,0%
SPECIALISM	STRING (3)	21	100,0%
HOSPITAL_DIAGNOSE	STRING (8)	2.078	100,0%
ICD10_DIAGNOSE	STRING (3-7)	3.298	66,0%
ICD10_PARENT	STRING (3)	1.161	66,0%
ICD10_BLOCK	STRING (7)	199	66,0%
ICD10_CHAPTER	STRING (6-10)	22	66,0%
HEALTHCARE_PRODUCT	INT (9)	3.395	100,0%
DECLARATION_CODE	STRING (6)	3.112	100,0%
DBC_COST	RAT (1-8)	2.847	99,9%

DBC_ID, EPISODE_ID, PATIENT_ID | Each row of the dataset represents a unique *DBC_ID* which represents a single patient diagnosis-treatment-combination. To differentiate DBCs for longer/chronic treatment plans alongside the *DBC_ID,* an *EPISODE_ID* is created to indicate if the patient arrives at the hospital for a new healthcare complaint or an existing one. To illustrate this with a practical example: if a patient visits the hospital again with the same complaint outside a predefined period a new *DBC_ID* is created, but the *EPISODE_ID* remains the same. If the same patient visits the hospital with a completely different healthcare demand, both a new *DBC_ID* and an *EPISODE_ID* are created. Every patient that visits the hospital is registered with a unique *PATIENT_ID*. More rows with the same *PATIENT_ID* indicate that a single patient has received multiple DBCs at the hospital. The *DBC_ID, EPISODE_ID,* and *PATIENT_ID* have been anonymised and labelled starting from the value 1000000, so that none could be translated back to a single patient.

GENDER | In the extracted dataset *GENDER* was indicated by M for male and V for female and has been translated to a binary variable. The gender of the patient in the final dataset is therefore indicated by a binary variable 1 for male, 0 for female.

STARTDATE_DBC, ENDDATE_DBC, BIRTH_DATE | *STARTDATE_DBC* provides the date when the DBC is opened, most often the date the patient arrives in the hospital for the first

time with a new healthcare complaint. *ENDDATE_DBC* provides the date when the DBC is closed, most often 42 days after an operation without new treatment for the same issue, 90 days after a regular treatment without another treatment for the same issue, or with a total maximum of 120 days. The *BIRTH_DATE* provides the date when the patient is born.

SPECIALISM, HOSPITAL_DIAGNOSE | Every DBC is specialism specific and is therefore categorized with a *SPECIALISM*, which is an abbreviation of three letters which corresponds to a three-digit code. The diagnose is represented by the *HOSPITAL_DIAGNOSE* and is a combination of the specialism three-digit code, a "-"sign and four digits/letters. Table 8 below gives more detail about the size and distribution of the different specialisms.

Table 8: Distribution of diagnoses per specialism

Hospital specialism	SPECIALISM	Specialism code	Unique hospital diagnoses	Unique number of DBC
Ophthalmology	OOG	301	74	222.515
Internal Medicine	INT	313	277	198.226
Cardiology	CAR	320	39	188.821
Surgery	HLK	303	209	179.574
Dermatology	DER	310	30	173.859
Orthopaedics	ORT	305	283	146.262
Paediatrics	KIN	316	62	124.944
Neurology	NEU	330	119	94.397
Gynaecology and Obstetrics	GYN	307	72	83.466
Urology	URO	306	78	72.918
Pulmonology	LON	322	49	72.457
Gastroenterology	GAS	318	70	67.444
Throat, Nose and Ear	KNO	302	249	48.727
Rheumatology	REU	324	78	31.003
Anaesthesia	ANA	389	33	24.295
Plastic Surgery	PCH	304	170	23.147
Geriatrics	GER	335	28	18.691
Rehabilitation	REV	327	47	16.626
Neurosurgery	NCH	308	19	5.856
Radiology	RAD	362	75	2.856
Psychiatry	PSY	329	17	1.383
Totals	**21**		**2.078**	**1.797.467**

ICD10_DIAGNOSE, ICD10_PARENT, ICD10_BLOCK, ICD10_CHAPTER **|** ICD10 is a medical classification list for disease diagnosis introduced by the WHO. The ICD10-standard is structured in a hierarchical framework which contributes to the classification of diagnosis and treatments. The exact conversion from *HOSPITAL_DIAGNOSE* to ICD10-standard has been known by the hospital since July 2015 and is actively in use since the start of 2020. Therefore, it is the only variable with missing data (34%), as previous data can't be fully labelled. The remaining 66% is a combination of original data and partially backpropagated labelling in order to convert as much *HOSPITAL_DIAGNOSE* into the ICD10-standard. By knowing the individual *ICD10_DIAGNOSE* the higher hierarchical levels are labelled via the structure of the ICD10 framework. To illustrate this structure the following practical example is given: an *ICD10_DIAGNOSE* (H25.2 Morgagnian lens type) belongs to an *ICD10_PARENT* (H25 Age-related cataract), belongs to the *ICD10_BLOCK* (H25-H28 Disorders of lens), which is part of the ICD10_CHAPTER (07 – VII Diseases of the eye).

HEALTHCARE_PRODUCT, DECLARATION_CODE, DBC_COST | Every patient that visits the hospital and receives a *HOSPITAL_DIAGNOSE* is also given a specific *HEALTHCARE_PRODUCT* and associated *DECLARATION_CODE* which are standardized on the national level. Multiple *HEALTHCARE_PRODUCTS* can have the same *DECLARATION_CODE,* which in the end is billed towards the insurance companies. Every *DECLARATION_CODE* has an accompanying average *DBC_COST* which is calculated yearly by VieCuri. The *DBC_COST* is not included in the original extracted dataset, but has been added through the linkage with *DECLARATION_CODE*. *DBC_COST* has a total of 0,1% missing data.

3.1.2 PRE-PROCESSING

In the previous section the selected variables which are being used as input for the method described in 3.2 are elaborated upon. Before being used as input for the method, the selected variables had to be pre-processed. The pre-processing for this research consists of two major parts, of which the first is the correct sequential mapping of the medical events for a single patient and the second is the aggregation of sequential representations of all patients in the dataset in order to serve as input vectors for the embedding of the machine learning algorithm.

The sequential mapping of a single patient consists of three smaller steps as is visualised in Figure 9. The first step is the mapping of a *PATIENT_ID* with all corresponding *DBC_ID*, ordered on their *STARTDATE_DBC*. As can be seen in the sequential patient mapping part of Figure 9 for patient 1000025 this resulted in 11 DBCs ordered from 22-8-2012 till 24-8-2018. Doing so creates a sequential mapping of all DBCs with their corresponding visit date for an individual patient. The second step is the mapping of the medical codes of interest, which for the visualisation in Figure 9 consists of the *HEALTHCARE_PRODUCT* variable. This mapping links the *DBC_ID* to a corresponding *HEALTHCARE_PRODUCT* again ordered sequentially on the *STARTDATE_DBC*. After doing so there are two mappings, both with the *DBC_ID* as unique identifier, ordered on *STARTDATE_DBC.* The last step then involves the mapping of the medical codes into the first mapping of the patients' visits in order to link the first and second step. This third steps enables the sequential ordering of the healthcare codes for a single patient.

After mapping every patient it is possible to represent a patient by a single row of data, where all medical codes are listed sequentially for that one patient. For patient 1000025 this would mean a row with eleven sequentially ordered medical codes as can be seen in Figure 9.

In order to aggregate all individual patients and store the data four datasets are generated. First a sequential listing of all unique *PATIENT_ID* starting from the integer 0 until 240.840, second a sequential listing of all unique *STARTDATE_DBC,* also stored as integers starting from 0 until 2.921 and third a listing of all *HEALTHCARE_PRODUCTS* also stored as integers starting from 0 until 3.394. These first three datasets function as the libraries of data as this has mapped all the linkages between integers and individual variables. The fourth and most important dataset, which also functions as the final input to the model, is the pickled or nested list of lists where the *HEALTHCARE_PRODUCT* variable inside the sequentially listed *DBC_ID* is sorted. Or in other words, the aggregation of all patient representations. A visualisation of this nested list of lists is shown in Figure 10. In Figure 10 every row represents a patient for which the *HEALTHCARE_PRODUCT* is listed in sequential order.

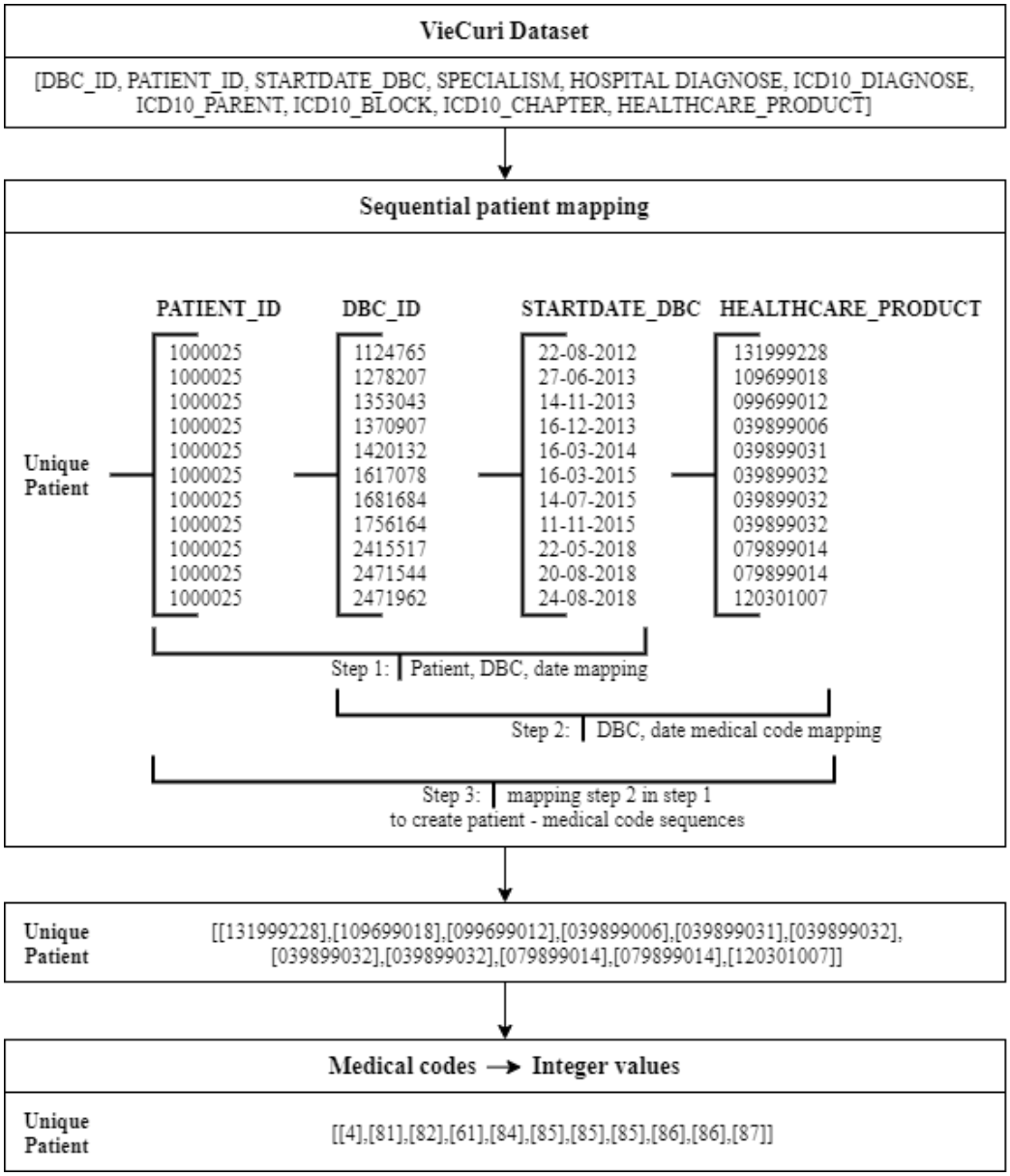

Figure 9: Mapping of single patient representation

"[[192], [79], [150], [154], [168], [52], [150], [52], [150], [52], [150], [150], [52], [150], [150]]"
"[[342], [343], [336], [215], [61], [344], [344], [344], [344], [345], [346], [263], [263], [263]]"
"[[326], [338], [326], [8], [246], [201], [217], [72], [10], [287], [347], [26], [26], [348], [290], [133], [134], [133]]"
"[[349], [350], [304], [86], [205], [204], [204], [325], [205], [246], [103], [351], [73], [103], [26], [79], [352], [133], [52], [79], [75], [105], [105], [52], [79]]"
"[[121], [353], [354], [355], [356], [357], [239], [92], [239], [239], [358], [359], [292], [194], [239], [292], [360], [121], [129], [361], [239], [194], [239], [292], [194], [292], [239], [362]]"
"[[259], [363], [334], [327], [110], [327], [328], [4]]"
"[[163], [364], [52], [365], [366], [52], [52], [169], [367], [257], [52], [257], [52], [257], [52]]"
"[[368], [97], [369], [97], [368], [97], [370], [64], [371], [64], [97], [372], [373], [286], [72]]"
"[[374], [76], [374], [91], [25]]"
"[[132], [86], [375], [376], [377], [139], [139]]"
"[[49], [364], [378], [378], [378], [378], [378], [46], [378], [378], [378]]"
"[[23], [0], [379], [380], [380], [380], [380], [380]]"
"[[33], [33], [33], [99], [334], [300]]"
"[[364], [132], [152], [19], [381], [132], [382], [383], [383], [384], [385], [386], [384], [386]]"
"[[79], [387]]"
"[[213], [91], [334], [334], [25], [34], [86], [334]]"
"[[388], [134], [167], [389], [133], [193], [390], [391], [390], [390], [390], [392], [390], [390], [393], [188], [269], [276], [390], [315], [192], [108], [315], [193], [193], [315], [276], [193], [315], [193]]"
"[[49], [394], [325], [325]]"

Figure 10: Aggregated sequential patient representations

As can be generally expected, and can be observed from Figure 10, the number of DBCs and thereby number of codes in the sequence differs for each patient. Not every patient visits the hospital with equal frequency and therefore the number of DBCs per patient varies. However, for any machine learning algorithm to process the data efficiently, the input shape of the data can't be of variable length. In order to overcome this issue the data is padded to the length of the longest sequence for that particular batch. The padding that is executed is the so-called zero padding at the end of the sequence. This entails that if within a batch the longest original sequence is 20, all smaller sequences become added zeros at the end, until the length corresponds with 20, in order to serve the machine learning algorithm with a single shape input.

For the purpose of predicting future medical events in a supervised setting, all patients who have less than two visits are removed from the dataset, as is done for the Doctor AI (Choi et al., 2016c) and MSAM (Zeng et al., 2020) researches. Patients with less than two visits have no present sequence and can therefore not be predicted. For the dataset at hand this entails the removal of 59.542 (24,7%) patients with only one single DBC within the timespan of January 2012 until December 2019. The total of 181.299 rows of patients with sequentially ordered codes that remain are than split in a 75% train, 15% test and 10% validation set.

3.2 DOCTOR AI ALGORITHM

Doctor AI, by Choi et al. (2016c), as introduced in the previous chapter, is a method for predicting future medical events of patients using the GRU architecture of RNN with a Skip-Gram approach of vector embedding. Doctor AI utilizes sequences of time and event pairs occurring in the patient's timeline as input to the GRU network. At every timestep the weight of a hidden unit is taken as the representation of the patient at that moment in time. From there it calculates and predicts future patient statuses accordingly. The method was originally tested on a real hospital dataset (260K patients over an eight year timespan) which in size is comparable to the one at hand (241K patients over an eight year timespan). On this dataset Doctor AI achieved 79,58% recall@30 after 20 training rounds on 1.183 diagnosis clusters. As explained by Al-Aiad et al. (2018) certain medical experts confirmed that Doctor AI was able to achieve human doctor level of predictive power and could provide meaningful clinical diagnosis prediction. In this paragraph first the embedding layer is explained in more detail, thereafter the RNN and GRU setup of Doctor AI are explained in 3.2.2, and the paragraph is closed with 3.2.3 which elaborates on the applied loss functions.

3.2.1 EMBEDDING LAYERS

First for every patient a multilevel point process is drawn for the observations in the form of (t_i, x_i) *for* $i = 1, \ldots, n$. Every pair indicates a visit in which medical codes are recorded in the DBC. At any time t_i, the multi-hot label vector $x_i \in \{0, 1\}^p$ indicates the provided healthcare product assigned, where p is the number of assigned codes. At a fixed interval t_i, higher-level codes can be extracted for predicting and are represented as y_i.

In order to improve the efficient representational learning of medical codes, the highest performing embedding variant used in Doctor AI, which incorporates Skip-Gram (Mikolov et al., 2013), was selected for this application. Skip-Gram has shown to be able to learn real practise multidimensional vectors to capture the latent representation of the medical codes

(Choi et al., 2016c). By incorporating this variant of embedding, it was able to lay out the product codes in a temporal order into the same lower dimensional space, so that similar related codes are embedded close to one another. This approach of initially learning the weight matrix (W_{emb}) between the input vectors and the embedding layer with the Skip-Gram algorithm allows the Doctor AI to only optimize the weights of the W_{emb} during training, instead of learning them from scratch. This way the weights of W_{emb} are improved as the whole model is trained. In short, this process allows the pre-training of the vectors in order to achieve better results. The formula which is used for this embedding approach is described below (1).

Multi-hot vector $$h_i^{(1)} = [x_i{}^{\mathrm{T}} W_{emb}, d_i] \quad (1)$$

In the last part of the embedding, multi-hot vectors, created by Skip-Gram, are converted to the vector representation, which is the input for the GRU/RNN setup, as explained in the following paragraph.

3.2.2 ARCHITECTURE OF THE RNN & GRU

After the creation of the multi-hot vectors in the embedding layer, the architecture of the RNN and GRU are of relevance. In Figure 11 below is demonstrated how the RNNs are implicated to predict the next healthcare product. In the diagram the first layer implants the higher dimensional input vectors in a lower-dimensional slot. The next layer has two sub layers, which depicts the recurrent units. This layer comprehends the status of the patient in each interval of data collection as a real-valued vector. By providing the status vector, two dense layers generate the codes observed in the next interval and the subsequent visit's schedule. In order to further explain the network, Figure 12 shows a visualisation of the GRU architecture for which the mathematical formula are given below all at timestep t_i.

Update gate $$z_i = \sigma(W_z x_i + U_z h_{i-1} + b_z) \quad (2)$$

Reset gate $$r_i = \sigma(W_r x_i + U_r h_{i-1} + b_r) \quad (3)$$

Intermediate memory unit $$\hat{h_i} = tanh(W_h x_i + r_i \circ U_h h_{i-1} + b_h) \quad (4)$$

Hidden layer $$h_i = z_i \circ h_{i-1} + (1 - z_i) \circ \hat{h_i} \quad (5)$$

In this setup of formulas the previous hidden layer (h_{i-1}) and current t_i input value x_i are the inputs for the update gate z_i (2) and reset gate r_i (3) as well as the intermediate memory unit $\hat{h_i}$ (4). Because of the σ in each gate functions the output value of these gates is between one and zero. If the reset gate represents a value close to zero, the intermediate memory unit will disregard the value from the previous hidden layer (h_{i-1}) and use the input x_i. If the output value of the update gate is close to one, the input value x_i is disregarded and the value of the previous hidden layer remains. Using this structure, the reset gate controls which information is disregarded and does not contribute to the prediction, whereas the update gate is responsible for the amount of information that is used from the previous hidden layer towards the current hidden layer (5).

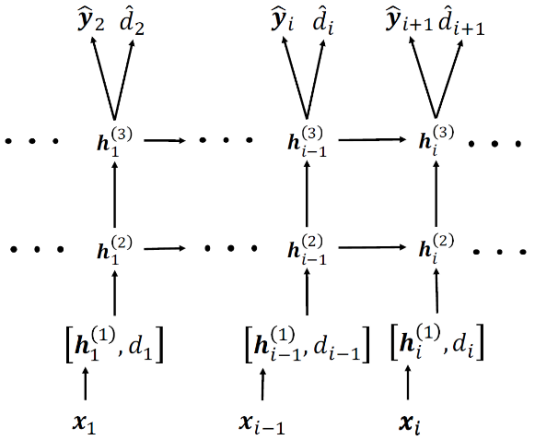

Figure 11: The RNN Doctor AI architecture (Choi et al. 2016c p. 5)

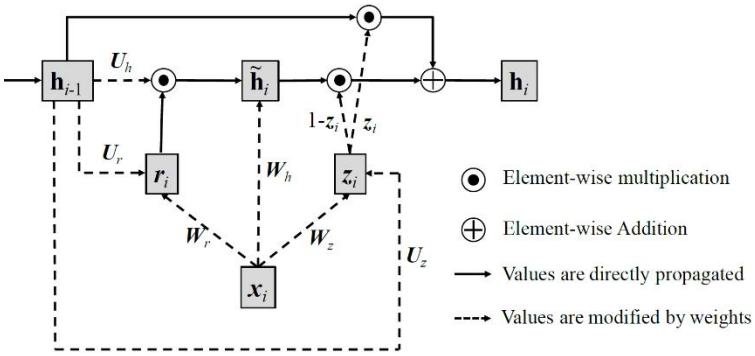

Figure 12: The GRU Doctor AI architecture (Choi et al. 2016c p. 14)

The goal of the Doctor AI algorithm is to comprehend a pragmatic presentation for the patient status at a fixed time interval t_i. The intention is to predict future healthcare product demand in the next visit y_{i+1}. In the end, all steps merge into a final single supervised learning scheme where the RNN architecture allows for learning the composed patient representations, improving the computation in the hidden layers using the patient's information to predict the next healthcare product in the sequence. The displayed architecture in Figure 11 takes input after a fixed interval t_i. The new inputs start from the engagement of multi-hot input vector x_i. After this, the lower-dimensional vector passes through the RNN's GRU setup as visualized in Figure 12. To improve the representation power of the network, the choice has been made to use two hidden layers, as was proposed in the best performing variant in the original method. The last step is the computation of the softmax loss function, which performs the prediction of the healthcare code and will be explained in the following paragraph.

3.2.3 LOSS FUNCTIONS

In order to perform medical code prediction at a fixed interval t_i, Doctor AI uses the softmax activation layer in combination with a cross-entropy loss function. In combination this variant is often called a softmax loss function, which in this case is a stacked layer on top of the GRU architecture. The softmax serves as activation function in order to normalize the output of the network over a probability distribution for the predicted output medical codes. In computing the softmax layer, Doctor AI uses the hidden layer h_i as input co calculate $\hat{y}_{i+1}$; the mathematical equation for this is displayed below (6).

The mathematical goal of training DoctorAI is to comprehend the weights $W_{\{z,r,h,code\}}$, $U_{\{z,r,h\}}$ and $b_{\{z,r,h,code\}}$. W's and U's values were initialized as ortho-normal matrices using singular value breakdown of matrices obtained by a normal distribution (Saxe et al., 2013), in contrast to the b's whose values started at zero. The total loss function is calculated using the final cross-entropy function. The cross-entropy function is used to sum up the negative logarithm of the probabilities of the softmax layer in order to quantify the difference between prediction and real-world values. The mathematical equation of the cross-entropy function as expressed for a single patient is also shown below (7).

Softmax function $$\hat{y}_{i+1} = softmax(W_{code}{}^{T} h_i + b_{code}) \quad (6)$$

Cross-entropy function $$\mathcal{L}(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{b}) = \sum_{i=1}^{n-1} \left\{ \left(\boldsymbol{y}_{i+1} \log(\hat{\boldsymbol{y}}_{i+1}) + (1 - \boldsymbol{y}_{i+1}) \log(1 - \hat{\boldsymbol{y}}_{i+1}) \right) \right\} \quad (7)$$

3.3 EVALUATION METRICS

Evaluating machine learning methods on their predictive performance is an essential part of every machine learning project. If not tested and evaluated against real-world datasets and metrics, the exercise is solely of academic relevance. In the case of Doctor AI (Choi et al., 2016c) the performance metrics of interest are those that give a clear understanding of the predictive power in a supervised multiclass classification task. Some performance metrics are more intuitive than others, however in order to explain them in more detail, first a general understanding of the in Figure 13 displayed confusion matrix within the medical context is needed.

The confusion matrix is a simplified representation of a binary classification task, which can be used to illustrate the underlying logic of the discussed evaluation metrics. In a confusion matrix as displayed below the correctly and incorrectly predicted instances are visualized compared to the actual value. For the problem at hand, which concerns multiclass classification, the number of classes (x) expands visually in size of x^2. Where in some cases one would be interested in optimizing the number of true positives (TP) and true negatives (TN) in order to achieve the highest accurate classification, in healthcare practice it can also be argued that incorrect predictions can have greater future healthcare risks, such as wrongly identifying a possible diagnosis. This problem arises when the potential cost of misclassification of a small class is high. A common example is dealing with a rare but health threatening disease. The cost of failing to diagnose (FN) the disease of an ill person is much higher than the cost of having wrongly predicted a disease (FP) and testing a healthy person on more metrics. Therefore, it is of importance to use evaluation metrics in healthcare prediction in a much broader perspective and with more precaution than for example predicting the electric consumption of a building.

	Predicted **YES**	Predicted **NO**
Actual **YES**	True Positive **TP**	False Negative **FN**
Actual **NO**	False Positive **FP**	True Negative **TN**

Figure 13: Confusion matrix 2x2 example

The issue which arises using common performance metrics as accuracy, precision and recall on large multiclass classification tasks is the high likelihood of class ambiguity. Class ambiguity arises when an algorithm has difficulties separating very similar classes. By increasing the number of classes, this problem becomes more relevant as whole groups of classes are more similar to one another than totally different. Within the healthcare domain this is even more prominent as some of the differences between groups of diagnoses and treatments are very nuanced when involving illnesses for a similar part of the human body or procedures for multiple purposes. Evaluating machine learning algorithm on a multiclass classification task by only taking into account the common performance metrics, would fail to capture the algorithms actual potential (Lapin et al., 2015; Lapin et al., 2016).

To tackle this issue the metrics which are most often used in evaluating machine learning methods within the multiclass classification and medical field are the so-called top-k metrics (k resembling an integer >=2). Top-k metrics differ as they do not only consider the highest probability in classifying an object, but consider the k highest probabilities. The simplest presentment of this principle is the ranking in which search engines present their results. In only considering the highest ranked link, one would not always find the answer to a given question, however when considering the top 2, 5 or 10 results, the chance of finding the right answer increases. Although the highest ranked link might have been on the right topic in the right context, it does not always resemble the correct option when comparing often 100.000 or more results. By looking at the highest k probabilities, the algorithm considers the class to be correct if it is within the top-k results.

The behaviour of the top-k metrics resembles that of a doctor conducting differential patient diagnoses. In this methodology the doctor lists the most probable diagnosis and moves along this list of options in order to treat the patient according to the identified diagnosis. Translated to our problem at hand, the doctor and therewith the top-k metrics would be able to apply the right treatment to the diagnosis, by systematically following the most probable solution to the patient's problem. The most commonly used evaluation metrics for methodologies similar to this research are either accuracy@top-k (Choi et al., 2017; Li et al., 2020a; Ma et al., 2017; Ma et al., 2018; Mu et al., 2018) or recall@top-k (Choi et al., 2016a; Choi et al., 2016c; Haq et al., 2017; Zeng et al., 2020). Accuracy@top-k is used to generates insight into the methods ability to correctly classify the output within the top-k results. Recall@top-k is used to generate insight into the methods ability to classify all relevant objects (FN) into the correct class and therewith compensates for the high medical cost involved for not identifying a relevant treatment in the top-k results. Machine learning methods, which can achieve both high accuracy@top-k and recall@top-k thus are able to mimic the doctor in applying the right healthcare product. The formulas which are used for both metrics are displayed below (formula 8 and 9).

$$\text{Accuracy@top-k} \qquad \frac{\text{number of correct predictions in top-k}}{\text{total number of predictions}} \tag{8}$$

$$\text{Recall@top-k} \qquad \frac{\text{number of true positive predictions in top-k}}{\text{number of total true positive predictions}} \tag{9}$$

4. EVALUATION OF RESULTS

In this chapter the performance of the Doctor AI (Choi et al., 2016c) algorithm is evaluated with the previously introduced evaluation metrics. Also, the potential financial impact of incorporating Doctor AI in the process of healthcare demand prediction is discussed. Therewith this chapter answers both the third and fourth sub-question of this research regarding the performance of the machine learning algorithm and the potential financial impact.

4.1 PERFORMANCE OF DOCTOR AI

For the translation from the methodology discussed in the previous chapter to an executable algorithm this research relies on input of other institutions, customised and developed to a final Python based code which fitted the VieCuri dataset. The original Doctor AI (Choi et al., 2016c) publicly released Python code heavily relied on the now outdated library Theano. Therefore, a later Python code approved by the authors built on the Pytorch platform, as published by The City College of New York (Russel-Puleri, 2019), is used as starting code for this research. However, for all pre-processing and sequential hot vector encoding the original Python code is used. All training has been done on a single Nvidia Tesla T4 GPU with 16GB of RAM memory, which took between 16 and 22 hours in order to execute the computation of 30 epochs of the Doctor AI algorithm for one single model configuration.

All different multiclass classification models which were executed varied in the number of predicted output classes (3.087 - 21) and number of nodes per hidden layer (200 or 400). The original Doctor AI used 200 nodes for each layer. The extra 400 node configuration is tested in every model to observe any differences in the yielded results. Aside from the original objective of predicting all healthcare products in the dataset, six other numbers of classes were tested to analyse the impact of different parameters and number of classes on the performance of the method. Beside the number of nodes per hidden layer, the hyperparameter settings were kept constant throughout the research. The settings for these hyperparameters are displayed in Table 9. For all applied models only patients with at least two visits were included. More detailed information on the differences between the models is shown in Table 10. Here the number of DBCs, patients, output classes, and random chance for the different models are listed. As this research is performed in a supervised setting, every model is only trained, validated, and tested on fully labelled data. Hence the difference in size of the data that is used for the different models. The random chance represents the expected success rate when patients were to be randomly assigned to a certain class.

Table 9: Doctor AI hyperparameter settings

Hyperparameter settings	
Embedding layer	1
Hidden layers	2
Nodes per hidden layer	200 or 400
Batch size	32
Learning rate	Adadelta (Zeiler, 2012) rate 0,01; rho 0,95
Dropout rate	0,5
Epochs	30

Table 10: Deployed Doctor AI models

Prediction	Number of classes	Number of DBCs	Number of patients	Random Chance
Healthcare Product (HP)	3.087	1.737.630	181.299	0,03%
HP 80% most sold	1.027	1.661.538	177.998	0,10%
HP 80% of total cost	519	1.245.028	153.481	0,19%
HP patients with >5 DBCs	3.081	1.533.119	108.002	0,03%
ICD10 Parent	1.132	1.132.913	140.961	0,09%
ICD10 Block	199	1.132.913	140.961	0,50%
Hospital Specialism	21	1.737.630	181.299	4,76%

The output classes which concern healthcare products (HP) are most relevant to the research, as they offer the highest level of detail on the expected healthcare demand. Together with the model for hospital specialism they are performed on the largest dataset, where patients were represented with at least two visits. Following this reasoning the data which only included patients with less than five visits or an 80% representation of the healthcare product are smaller in size. The results on both accuracy@k and recall@k which were retrieved from the training of the healthcare products model of the Doctor AI algorithm are presented in Table 11. In the table the highest percentage for both performance metrics is marked per model. All considered performances are from data which were new to the model (holdout/test set 15%). All deployed models in this research showed a slightly bigger cross entropy value on the test set in comparison to the validation set, but decreased around the same slope.

Table 11 clearly shows that the performance of predicting all 3.087 healthcare products simultaneously is very low on both metrics. With a maximum value of 0,76% accuracy@30 and 1,54% recall@30 the outputs of the presented model show to be an improvement over random chance (0,03%), but come nowhere near a good predictive model or the 79% recall@30 from the original Doctor AI research. From Figure 17 in Appendix C can be derived that the 200 nodes per hidden layer configuration shows a climbing incremental convergence as epochs progress. However, the 400 nodes per hidden layer suffers a great decrease in performance after the 14th epochs. As it shows this behaviour in multiple attempts for this model, it is expected that either the unbalance of classes within the batches is too high, the gradient decent is stuck in a sharp minimum, or the model is overfitting as it converges too fast. All three arguments could be contributing to the decrease in performance, however due to the observed fluctuation in the 400 nodes curve in Figure 17 and the much sharper decline in test loss, the conclusion of overfitting is most prominent.

Table 11: Performance metrics for healthcare product (HP) model (n=3.087)

Prediction	Nodes	Epoch	Accuracy@top-k			Recall@top-k		
			Acc@10	Acc@20	Acc@30	Rec@10	Rec@20	Rec@30
Healthcare Product (HP) n=3.087	200	10	0,25%	0,51%	0,76%	0,83%	0,92%	1,12%
		20	0,25%	0,51%	0,76%	1,01%	1,13%	1,45%
		30	0,25%	0,51%	**0,76%**	1,10%	1,21%	**1,54%**
	400	10	0,25%	0,51%	0,76%	0,81%	0,94%	**1,27%**
		20	0,25%	0,51%	0,76%	0,22%	0,30%	0,39%
		30	0,25%	0,51%	**0,76%**	0,25%	0,33%	0,37%

As the model on healthcare product prediction did not show good results, the six other models have been tested. In general it can be concluded that the setup of 400 or 200 nodes is not superior to the other in every situation, but depending on the specific application, one can outperform the other. However, it is clear that the model setup with 400 nodes per hidden layer shows faster conversion speed than that with 200 nodes. This phenomenon in itself is not strange as more nodes can essentially "store" more granular value than less nodes.

The two alternative models which are tested first are both focused on the healthcare products output, while decreasing the number of output classes. As described in section 1.1 over the years 2017 to 2019 a total of 1.027 products were responsible for 80% of the total sold products. However, when looking at the products which are responsible for 80% of the total cost, the mix is decreased to 519 different products. Both model configurations are trained, validated and tested by only incorporating the relevant 1.027 or 519 classes and thereby leaving out all rare and low total impact on cost products. However, leaving out all the small groups of healthcare products, as derived from Table 12 does not show any big increase in performance. The accuracy@30 for both models increases from 0,76% to 2,29% in the best configuration for the most sold products model and to 4,53% for the highest contribution to cost model. Whereas the recall@30 shows a decrease for the most sold products and only a slight increase from 1,54% to 2,42% for the products responsible for 80% of cost.

Table 12: Performance metrics for subsets of HP models (n=1.027 and n=519)

Prediction	Nodes	Epoch	Accuracy@top-k			Recall@top-k		
			Acc@10	Acc@20	Acc@30	Rec@10	Rec@20	Rec@30
80% Most Sold HP n=1.027	200	10	0,76%	1,53%	2,29%	0,04%	0,13%	0,17%
		20	0,76%	1,53%	2,29%	0,03%	0,10%	0,16%
		30	0,76%	1,53%	**2,29%**	0,03%	0,11%	**0,18%**
	400	10	0,76%	1,52%	2,29%	0,02%	0,09%	0,15%
		20	0,76%	1,52%	2,29%	0,05%	0,12%	0,40%
		30	0,77%	1,53%	**2,29%**	0,41%	0,46%	**1,04%**
80% Of Total HP Cost n=519	200	10	1,51%	3,02%	4,53%	0,00%	0,39%	0,39%
		20	1,51%	3,02%	4,53%	0,13%	0,91%	0,92%
		30	1,51%	3,02%	**4,53%**	0,20%	0,99%	**0,99%**
	400	10	1,51%	3,01%	4,52%	0,04%	1,01%	1,02%
		20	1,51%	3,02%	4,52%	0,85%	2,17%	2,20%
		30	1,51%	3,02%	**4,53%**	0,75%	2,37%	**2,42%**

The results of Table 12 show that by significantly decreasing the number of output classes while remaining the same level of detail, results only show a significant increase in the accuracy@top-k metric. This means that a decrease in number of output classes, while keeping the most frequent occurring healthcare products, increases the algorithms capabilities of rightfully predicting the correct class. On the other hand, it also shows that in doing so, the model in general does not improve in the recall@top-k performance, which indicates a worse performance in recognising false negatives. Only the 400 nodes model of the 80% of total healthcare product cost shows an improved performance (recall@30, 2,42%) in better detecting false negatives. In all other recall@top-k metrics the models show lower performance when compared to the original 3.087 classes healthcare product model. Although the combined

accuracy@top-k and recall@top-k is slightly better than the original healthcare product model, the performance is still very low, when considering that the number of classes have been reduced greatly in both models. The performance visualisation can be found in Appendix C (Figure 18 and Figure 19).

The next model that is tested is a setup which yielded better results in the Dipole (Ma et al., 2017) methodology by only incorporating patients with at least five recordings. This model as derived from Table 13 below and Figure 20 in Appendix C did not improve the performance metric results on the VieCuri dataset. Both the 200 nodes per hidden layer setup with accuracy@30 (0,70%) and recall@30 (0,78%), as well as the 400 nodes setup with accuracy@30 (0,70%) and recall@30 (0,40%) performed worse than the original model. In the configuration of the Dipole (Ma et al., 2017) method, the researchers assumed that patients with less than five recordings were not frequent enough visitors to include in the dataset. For their dataset this meant that less than 2,5 visits per patient per year were not incorporated. In this model, with the VieCuri dataset, the density for the visits per patient per year went up from 1,06 to 2,03. However, the exclusion of all patients with 2-4 recordings resulted in a decrease in performance as can be concluded from this model. Probable cause for these worse performance metrics is the loss of almost 40% of patients and thereby losing too many sequential relevant relationships, which were not compensated for by the more dense remaining patients.

Table 13: Performance metrics for >5 DBCs model (n=3.081)

Prediction	Nodes	Epoch	Accuracy@top-k			Recall@top-k		
			Acc@10	Acc@20	Acc@30	Rec@10	Rec@20	Rec@30
HP Patients With >5 DBCs n=3.081	200	10	0,23%	0,47%	0,70%	0,26%	0,61%	0,66%
		20	0,23%	0,47%	0,70%	0,31%	0,57%	**0,78%**
		30	024%	0,47%	**0,70%**	0,36%	0,55%	0,71%
	400	10	0,23%	0,47%	0,70%	0,13%	0,23%	0,29%
		20	0,23%	0,47%	0,70%	0,17%	0,28%	0,38%
		30	024%	0,47%	**0,70%**	0,17%	0,30%	**0,40%**

The models which are widely used, and can be found in Doctor AI (Choi et al., 2016c), Dipole (Ma et al., 2017) and MSAM (Zeng et al., 2020) are the clustering of groups of codes into a higher hierarchical level. For the purpose of this research the ICD10-framework which the VieCuri registers is used in order to maintain (part of) the medical diversity, while still decreasing the number of output classes significantly. The prediction of ICD10-parents instead of individual codes and the blocks of medical groups the codes belong to are, in comparison to healthcare products, a big decrease in the number of predicted classes (1.132 parents and 199 blocks). The model predicts in which cluster of treatments a healthcare product is situated and presents therewith a higher hierarchical level of the medical healthcare product codes. It is therefore no surprise that the results in Table 14 show improved maximum values of 2,07% accuracy@30 and 8,18% recall@30 for the ICD10-parent prediction and maximum values of 11,80% accurarcy@30 and 27,62% recall@30 for the ICD10-block prediction.

Table 14: Performance metrics for ICD10 models (n=1.132 and n=199)

Prediction	Nodes	Epoch	Accuracy@top-k			Recall@top-k		
			Acc@10	Acc@20	Acc@30	Rec@10	Rec@20	Rec@30
ICD10 Parent n=1.132	200	10	0,69%	1,38%	**2,07%**	0,22%	2,84%	2,87%
		20	0,69%	1,38%	2,07%	0,23%	6,15%	6,18%
		30	0,69%	1,38%	2,07%	0,13%	6,69%	**6,70%**
	400	10	0,69%	1,38%	2,07%	2,77%	7,82%	7,84%
		20	0,69%	1,38%	**2,07%**	2,34%	8,16%	**8,18%**
		30	0,69%	1,38%	2,07%	2,02%	7,92%	7,93%
ICD10 Block n=199	200	10	3,92%	7,85%	11,77%	1,44%	14,68%	17,51%
		20	3,93%	7,86%	**11,80%**	2,12%	20,09%	25,55%
		30	3,93%	7,85%	11,78%	1,85%	19,86%	**26,02%**
	400	10	3,93%	7,86%	11,79%	2,64%	21,13%	**27,62%**
		20	3,93%	7,86%	11,78%	1,74%	20,27%	27,16%
		30	3,93%	7,86%	**11,80%**	1,71%	19,97%	26,44%

The adoption of a higher hierarchical level contributes to better predictive outcomes on fewer classes. In Figure 21 and Figure 22 in Appendix C can be seen that the setup of 400 nodes per hidden layer in both models has a much faster rate of convergence than the 200 nodes setup. The 200 nodes setup shows a more slowly progressing increase whereas the 400 nodes model starts fluctuating, which is an indication of overfitting. Although an overall improvement in the predictive performance of the model can be observed, the ICD10-parent and -block do not achieve high enough levels to be incorporated in any forecasting praxis. Both models lose the more detailed information of the healthcare products, but still contain the information which is used to indicate the severity and magnitude of the expected treatment.

4.1.1 HOSPITAL SPECIALISM MODEL

The best performing and highest hierarchical level tested in this research is the model that classifies the hospital specialism. VieCuri hosts 21 different specialisms, which offer the 3.087 different healthcare products. Therefore, this model represents a clustering of the healthcare products on a specialism level. In predicting the hospital specialism the finer information on product or treatment cluster are lost, however it shows where the patients will receive their next treatment. Although more difficult in allocating exact costs, this level offers insight into the number of patients per specialism and can therefore still be of value to the hospital. However, in reducing the output classes to 21, using the same number of top-k would not represent adequate evaluation metrics as @10 would represent already 47,6% of classes and @30 more than 100%. Therefore, with the scale down from 199 output classes in the ICD10-block prediction model to the 21 in the hospital specialism, the top-k metrics have been decreased in the same order of magnitude towards @1 (which would represent the common/none top-k metric), @2 and @3 for both accuracy@top-k and recall@top-k. It is no surprise that due to the reduction of output classes in this model and the same volume of data, this model is the best performing overall model with maximum achieved values for both accuracy@3 of 75,36% and recall@3 of 72,64% for the 400 nodes per hidden layer setup (Table 15). The 200 nodes model also performs well with accuracy@3 of 74,42% and recall@3 of 68,06%.

Table 15: Performance metrics for hospital specialism model (n=21)

Prediction	Nodes	Epoch	Accuracy@top-k			Recall@top-k		
			Accuracy	Acc@2	Acc@3	Recall	Rec@2	Rec@3
Hospital Specialism n=21	200	10	25,87%	61,77%	74,41%	23,63%	44,98%	63,68%
		20	25,87%	61,72%	74,41%	25,72%	47,28%	66,67%
		30	25,88%	61,77%	**74,42%**	26,57%	51,69%	**68,06%**
	400	10	25,86%	61,74%	75,30%	26,17%	50,70%	69,07%
		20	25,88%	61,77%	75,31%	27,98%	53,49%	71,86%
		30	25,88%	61,78%	**75,36%**	28,04%	54,27%	**72,64%**

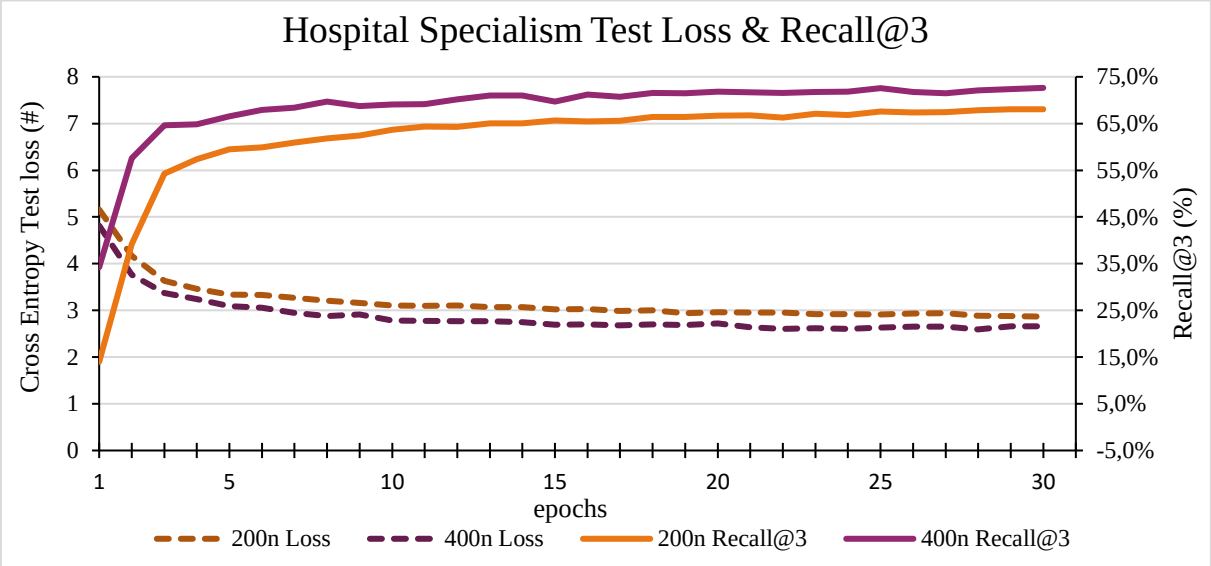

Figure 14: Loss and recall@3 for hospital specialism model (n=21)

Similar to most of the other presented models, the hospital specialism model shows a clear gradient conversion towards the maximum recall@3 values at epochs 30 in Figure 14. As can be derived from the dotted lines in Figure 14, the 400 nodes per hidden layer model achieves a slightly faster decrease in total cross entropy loss compared to the 200 nodes; therefore, it is no surprise that the observed recall@3 has a faster conversion towards the maximum value for the 400 nodes model. Overall, both the 400 nodes per hidden layer and 200 nodes model show the same behaviour in both gain in recall@3 and decrease in total loss. As the hospital specialism model is the best performing model under consideration, a normalized 21x21 confusion matrix for the 400 nodes per hidden layer model is compiled in Figure 15.

The confusion matrix shows the difference in performance for the VieCuri specialism with a false positive and false negative tendency towards the classes with the highest occurrence in the dataset. In Figure 15 the highest probabilities are marked orange whereas the lowest probabilities are purple. With the exception of the Anaesthesiology and Geriatrics specialisms (in the matrix displayed as ANA and GER) it can be concluded that the lowest represented classes overall are the weakest performers, whereas an increase in the number of predicted classes also improves the overall performance. This behaviour confirms the reasoning of Choi et al. (2016c) that the rarity of medical codes highly influences the model's performance and is a result of working with diverse real-world patient data (Ruan et al., 2019; Sheets et al., 2017). The general insight which can be derived from the confusion matrix is that the algorithm is capable of distinguishing the different classes.

		Predicted classes																		
		CAR	HLK	INT	ORT	LON	OOG	GAS	GYN	URO	NEU	KNO	KIN	DER	PCH	GER	REU	ANA	NCH	REV
True classes	CAR	0,57	0,07	0,06	0,07	0,03	0,02	0,02	0,00	0,03	0,02	0,02	0,02	0,02	0,00	0,00	0,01	0,00	0,02	0,03
	HLK	0,12	0,64	0,04	0,03	0,02	0,03	0,01	0,01	0,03	0,01	0,01	0,00	0,02	0,00	0,02	0,01	0,00	0,00	0,00
	INT	0,03	0,03	0,73	0,08	0,02	0,04	0,00	0,00	0,01	0,00	0,04	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00
	ORT	0,05	0,01	0,09	0,58	0,02	0,04	0,03	0,02	0,01	0,03	0,05	0,03	0,03	0,00	0,00	0,00	0,00	0,00	0,00
	LON	0,06	0,14	0,02	0,07	0,51	0,03	0,04	0,01	0,01	0,00	0,07	0,00	0,03	0,00	0,00	0,00	0,00	0,00	0,00
	OOG	0,07	0,06	0,03	0,01	0,02	0,73	0,03	0,01	0,00	0,02	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00
	GAS	0,09	0,05	0,01	0,04	0,02	0,00	0,54	0,05	0,03	0,01	0,02	0,04	0,05	0,00	0,02	0,03	0,00	0,00	0,00
	GYN	0,08	0,06	0,04	0,05	0,09	0,05	0,03	0,46	0,04	0,04	0,03	0,01	0,00	0,01	0,00	0,02	0,00	0,00	0,00
	URO	0,03	0,03	0,08	0,04	0,00	0,05	0,00	0,03	0,60	0,00	0,02	0,01	0,02	0,00	0,03	0,05	0,00	0,00	0,00
	NEU	0,04	0,02	0,00	0,02	0,00	0,07	0,03	0,11	0,00	0,64	0,00	0,01	0,04	0,00	0,00	0,00	0,00	0,00	0,00
	KNO	0,04	0,03	0,00	0,08	0,02	0,06	0,00	0,02	0,01	0,05	0,64	0,00	0,03	0,00	0,00	0,01	0,00	0,00	0,00
	KIN	0,05	0,07	0,04	0,08	0,04	0,03	0,00	0,00	0,02	0,00	0,00	0,62	0,03	0,00	0,00	0,01	0,00	0,00	0,00
	DER	0,00	0,02	0,03	0,01	0,01	0,05	0,02	0,01	0,02	0,03	0,06	0,01	0,69	0,01	0,01	0,02	0,01	0,00	0,00
	PCH	0,08	0,17	0,00	0,00	0,00	0,03	0,01	0,00	0,11	0,00	0,02	0,00	0,00	0,55	0,00	0,02	0,00	0,00	0,00
	GER	0,19	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,79	0,00	0,00	0,00	0,00
	REU	0,12	0,00	0,00	0,03	0,00	0,04	0,00	0,04	0,05	0,09	0,17	0,00	0,00	0,11	0,00	0,36	0,00	0,00	0,00
	ANA	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,00	0,90	0,03	0,00
	NCH	0,00	0,13	0,05	0,07	0,01	0,20	0,00	0,04	0,00	0,00	0,06	0,06	0,00	0,04	0,00	0,07	0,00	0,27	0,00
	REV	0,09	0,22	0,04	0,00	0,05	0,08	0,00	0,05	0,00	0,04	0,00	0,03	0,03	0,00	0,00	0,00	0,06	0,00	0,32
Total		2.475	2.636	2.345	2.585	1.038	3.299	1.127	1.528	1.245	1.364	2.153	901	2.753	428	277	596	305	69	72

Figure 15: Confusion matrix 21x21 for hospital specialism model (400 nodes)

Altogether, the results from the Doctor AI (Choi et al., 2016c) multiclass classification implementation show low performance in predicting a high number of output classes simultaneously on the VieCuri dataset in contrast to the original research. By moving to a higher hierarchical level, and therewith decreasing the number of output classes, the performance of the models significantly increases. The other efforts in focussing on most sold, most total cost contributors, or patients with more than five visits did not improve performance, as a lot of visit relationships were removed. From the observed behaviour of the cross entropy loss and recall@top-k functions of the different models it can be concluded that the 400 nodes per hidden layer configuration of the different models show faster learning speed. However, they do not always yield overall better results in contrast to the 200 nodes per hidden layer from the original setup. Prediction of healthcare products and ICD10-parents contains the most valuable information to the VieCuri organisation, but did not perform anywhere close to the original model, or well enough to be deployed in prediction practice. Only the model with the most generalizable results and the least output classes remains for incorporating any of the above achieved results into the VieCuri healthcare forecasting praxis. The hospital specialism model (with 400 nodes per hidden layer) which is also visualised in the confusion matrix is used in the business case presented in the following paragraph.

4.2 POTENTIAL FINANCIAL IMPACT

The fourth and last sub-question is concerned with the potential financial impact of incorporating machine learning in the yearly forecasting of healthcare demand for VieCuri Medisch Centrum. In order to access any potential financial impact of incorporating the Doctor AI (Choi et al., 2016c) algorithm in the yearly forecasting, the situation for the last three years of data is observed. As the healthcare product and ICD10-model configurations did not yield high enough results, only the best performing model (the 400 nodes per hidden layer) of the hospital specialism model is extrapolated from the number of predicted instances to the whole dataset. As the structured extraction of financial data was limited to the three financial years of 2017 to 2019, the extrapolation of financial impact is also calculated for these three years. More details on the ratio between the number of DBC and incorporated number of DBC in the predicted scenario are displayed in Table 16.

Table 16: Number of total and predicted DBCs and associated cost

	2017		**2018**		**2019**	
Total DBCs	234.904	100%	248.807	100%	254.011	100%
Last in sequence DBCs	18.917	8,1%	28.885	11,6%	88.445	34,8%
Test set DBCs	3.216	1,3%	4.188	1,7%	13.090	5,2%
Total DBC cost	€ 174.042.005	100%	€ 180.929.072	100%	€ 182.128.328	100%
Last in sequence DBC cost	€ 11.311.272	6,5%	€ 15.832.476	8,8%	€ 46.148.690	25,3%
Test set DBC cost	€ 1.994.944	1,1%	€ 2.445.210	1,4%	€ 7.688.087	4,2%

Doctor AI (Choi et al., 2016c) focused on predicting the last code within the patient sequence, which means not all occurred codes for patients within 2017 to 2019 were predicted in the final model. As can been seen from Table 16 for the year 2017, there were a total of 234.904 DBCs included in the VieCuri dataset, and of those 18.917 turned out to be the last DBC in the patient sequences. However, the dataset was split into a 75% train, 10% validation, and 15% test set. Therefore, of the 18.917 last in sequence DBC, a total of 3.126 were represented in the test set of which the results are reported in the paragraph above. The difference between the total amount of DBCs sold to the insurance companies and the number of DBCs present in the last class of the patient sequences results in the extrapolation for a large group of patients. The maximum amount (88.445) of available predictions in the last class of the sequence only accounts for 34,8% of the total number of DBCs for the year 2019. As the average visit per patient per year in the VieCuri dataset is 1,06, it is not strange that most final visits occurred in the last three years of data for an average patient. Since in healthcare practice there is no such thing as an average patient, not all final patient visits are within the last three years of data. A patient who only visited the hospital twice and did so in 2012 and 2013 for example, is included in the performance of the model in section 4.1, but is disregarded as having financial impact on the period thereafter. As such, the number of predicted patients that is used in the extrapolation of the total costs and occurred in 2017 to 2019 accounts for 75,2% of all predicted 181.299 patients' final code in the sequence. This means that for 24,8% of the included patients the sequence of treatment at VieCuri ended between 2012 to 2016.

The model which predicted the hospital specialism (400 nodes per hidden layer) is used in order to calculate the financial impact in terms of fault margin and loss of treatment compensation over the years 2017 to 2019. The distribution of DBCs that is used to extrapolate the Doctor AI results is the same distribution as is shown in the confusion matrix of Figure 15. The healthcare product costs within a specialism are averaged for the particular year in order to multiply the number of identified patients to be treated at the specific specialism. Thus, the calculations displayed in Table 17 do not show under and over estimations on healthcare product level, but concentrate on the overall difference within a hospital specialism. The calculation for the loss of treatment compensation and margin of error remains the same as those of section 1.1. Therefore, the two specialisms Radiology and Psychology (which together contribute to 0,002% of DBC total cost) are divided amongst the other 19 specialisms which ordered the execution within the domain. This way Table 17 generates a one-on-one comparison between the current VieCuri practice from section 1.1 and the extrapolation of results of the hospital specialism model.

Table 17: Financial impact forecasting performance per specialism

Hospital specialism		**2017**		**2018**		**2019**	
		VieCuri	**Model**	**VieCuri**	**Model**	**VieCuri**	**Model**
Cardiology	CAR	-0,2%	-1,6%	-1,7%	5,8%	3,6%	7,6%
Surgery	HLK	0,9%	14,0%	2,7%	11,1%	-3,6%	1,2%
Internal Medicine	INT	1,7%	-20,0%	2,6%	-1,3%	0,6%	17,3%
Orthopaedics	ORT	8,7%	-1,2%	2,2%	-19,0%	0,8%	10,3%
Pulmonology	LON	2,0%	-4,7%	-0,1%	-14,7%	1,5%	12,3%
Ophthalmology	OOG	1,5%	-3,3%	4,8%	9,5%	6,9%	21,4%
Gastroenterology	GAS	2,3%	-0,8%	11,7%	15,4%	12,3%	10,5%
Gynaecology and Obstetrics	GYN	7,3%	63,3%	1,4%	86,2%	-2,5%	80,8%
Urology	URO	4,9%	-24,0%	7,3%	-18,5%	12,6%	13,2%
Neurology	NEU	1,9%	9,9%	9,2%	16,5%	0,7%	-7,8%
Throat, Nose and Ear	KNO	2,1%	42,9%	1,9%	13,4%	4,3%	30,6%
Paediatrics	KIN	7,4%	16,7%	-8,0%	8,8%	-7,7%	-3,1%
Dermatology	DER	-2,4%	22,6%	3,8%	30,0%	-2,1%	46,0%
Plastic Surgery	PCH	7,4%	23,5%	10,3%	34,4%	15,7%	58,1%
Geriatrics	GER	22,2%	28,8%	4,1%	-4,8%	-14,1%	17,2%
Rheumatology	REU	1,9%	-15,5%	12,8%	-7,1%	-0,9%	34,5%
Anaesthesiology	ANA	-6,4%	-1,5%	18,4%	4,1%	19,1%	2,6%
Neurosurgery	NCH	13,6%	41,7%	-5,9%	53,7%	-12,2%	16,2%
Rehabilitation	REV	-0,4%	35,1%	0,1%	0,1%	0,1%	-0,1%
Loss of treatment compensation VieCuri Forecasting €		**€ 5.129.930**		**€ 8.065.545**		**€ 8.855.869**	
Average fault margin VieCuri Forecasting %		**3,0%**		**4,8%**		**5,2%**	
Loss of treatment compensation Extrapolated DoctorAI €		**€ 23.797.610**		**€ 25.537.184**		**€ 27.315.252**	
Average fault margin Extrapolated DoctorAI %		**14,0%**		**15,0%**		**16,1%**	

From Table 17 can be concluded that extrapolating on the hospital specialism prediction does not perform better in forecasting healthcare demand for VieCuri. Using this high-level information, a distinction cannot be made between surgical operation and more simple treatments or treatment which requires multiple admission days and one day treatments. As healthcare products move within a wide range of prices (€5-€65.000), this method assumes great generalizability between patient treatment. Although all within the correct order of magnitude, the extrapolated results from those predictions are therefore only a rough estimate. It comes as no surprise that the loss of treatment compensation and average fault margin by extrapolating the hospital specialism results for all three years: €23.797.610 (14,0%), €25.537.184 (15,0%), and 27.315.252 (16,1%) are bigger when compared to the current VieCuri practise. Although some individual hospital specialism extrapolations show a fault margin that is not far off, as for example the year 2019 for Surgery (1,2%) and 2018 for Internal Medicine (-1,3%). Over all three years it may be concluded that, with the exception of Anaesthesiology (-1,5%, 4,1% and 2,6% fault margins by extrapolating), all fault margins are worse than the current VieCuri practise. The biggest outlier, which becomes clear from Table 17, is the Gynaecology specialism. The specialism recognises a great spike in patients who are born, which typically is the first encounter with a healthcare product and accounts for 53,6% of the budget and 29,1% of total treatments. Therefore, logically this specialism represented the biggest fault margin, as the algorithm greatly relies on historical sequenced data in order to predict the next visit, a mechanism which does not predict the first visit well.

5. CONCLUSION AND DISCUSSION

The final chapter of this research is focussed on answering the overall research question by integrating the information gained the previous chapters. Subsequently in the conclusion three paragraphs are dedicated to going more in depth on the practical and theoretical implications as well as the limitations of this research.

5.1 CONCLUSION

This research aims to answer the question whether machine learning is able to predict healthcare product demand more accurately than the VieCuri current practise. For practical purposes hospitals can use these insights to better predict and manage the budgeting of specific treatments, obtain a more accurate forecast and keep treatment costs affordable. For theoretical purposes researchers gain insights in the behaviour of machine learning methods on different datasets and the generalisability of performance. Therefore, the main research question that is asked at the beginning of this research is:

How can VieCuri Medisch Centrum more accurately forecast healthcare demand applying machine learning on EHR data in order to decrease loss of treatment compensation?

In order to answer the main research question four sub-questions were developed, as elaborated upon in section 1.2. Together with the CRISP-DM framework (Shearer, 2000) the four sub-questions have functioned as the phasing throughout this research. Based on the literature review conducted for this research the Doctor AI (Choi et al., 2016c) machine learning algorithm was selected as most prominent and practical methodology for the problem of forecasting healthcare demand. In parallel the EHR dataset from the VieCuri Medisch Centrum data warehouse was extracted in order to validate and test the performance of Doctor AI (Choi et al., 2016c) on the local hospital dataset. Subsequently the performance was extrapolated to show the financial impact of this machine learning algorithm on the yearly healthcare demand forecasting methodology in VieCuri.

Dutch hospitals are challenged with providing quality healthcare to an aging population in a cost-efficient manner. Hospitals that effectively leverage machine learning capabilities in forecasting healthcare demand can potentially capture more insurance compensation. Although more profit does not always directly contribute to better healthcare, generating yearly positive profit margins is generally considered necessary to support sustainable high healthcare quality (Beauvais & Wells, 2006). Capturing more insurance compensation by forecasting more accurately can enable hospitals to invest in important areas such as the training of employees, expanding on capacity, funding research, and keeping healthcare affordable.

From the conducted literature study can be concluded that although different variations and applications of machine learning are being developed recently, the underlying dominant approach within the healthcare demand prediction is deep learning. Although each encountered deep learning methodology has its unique features, the underlying strength of detecting hidden patterns in data and leveraging these to predict scalable end-to-end algorithms is prominent and necessary to process large amounts of EHR data. Acknowledged by many of the incorporated researches (Galatzer-Levy et al., 2014; Jiang et al., 2017; Khaldi et al., 2017; Miotto et al., 2016; Prasad & Agarwal, 2014; Roysden & Wright, 2015; Srinivas et al., 2010; Yang et al.,

2017; Zlotnik et al., 2015) more validation of methods which rely on data from a single source or geographic region should be done in order to show generalizable forecasting power on other EHR datasets before qualifying them as good forecasting methods. By researching and deploying the Doctor AI (Choi et al., 2016c) as one of the most promising identified methods, this thesis attempts to contribute to the validation of machine learning practise, and addresses the healthcare demand forecasting of VieCuri. The RNN/GRU architecture of Doctor AI shows a literature backed track record of generalizable and broad predictive power. However, when trained and validated on the VieCuri EHR dataset, the algorithm does not perform as expected. The method is tested on different models with a variety of output classes and hierarchical levels and in general yields lower results than needed to predict on future healthcare product demand or treatment clusters (by ICD10-standard). The model configuration which did yield satisfying results was the multiclass classification on the 21 different specialisms present in the VieCuri hospital. This model achieves 75,4% accuracy@3 and 72,6% recall@3 and is thereby the candidate which is used to show the potential financial impact of incorporating machine learning in healthcare demand forecasting.

In comparing the current forecasting fault margins per specialism with the extrapolated predicted hospital specialism from Doctor AI, it becomes clear that information on the differences in cost for the individual healthcare products is lost in the Doctor AI model of hospital specialism. Leveraging average cost per patient visit is a compromise between missing more detailed cost versus high enough predictive power of the algorithm to be of use. The fault margins by applying this method are 14,0% for the year 2017, 15,1% for the year 2018, and 16,1% for the year 2019. Although in the same order of magnitude, this variation does not yield better performance in predicting yearly future healthcare demand than the current VieCuri praxis which yields fault margins of 3,0% (2017), 4,8% (2018), and 5,2% (2019).

It can be concluded that machine learning in general has huge potential in predicting future demand as well as handling the unique and challenging nature of EHR data. On the other hand, this research illustrates that not every algorithm achieves the same level of performance when applied to a different dataset in a different setting. The results of this research show that the applied machine learning method in combination with the nature of the available EHR data is capable of predicting the high-level hospital specialism. However, it fails to achieve satisfying results when the hierarchical level is decreased and therewith the number of output classes increases. Furthermore, it shows that with the performance of Doctor AI (Choi et al., 2016c) on the VieCuri dataset it is not possible to decrease the prediction fault margin and loss of treatment compensation by applying it to healthcare demand forecasting praxis.

5.2 LIMITATIONS AND FUTURE RESEARCH

This research has several limitations which can be directions for future research. In this section the three most prominent limitations will be further elaborated upon. The first limitation of this research can be recognised in both time and capabilities of the researcher. In the best case scenario a research within the field of machine learning application would test and compare a variety of methods to come up with the best performing algorithm. Besides the implemented Doctor AI (Choi et al, 2016c), the two other identified methods Dipole (Ma et al., 2017) and MSAM (Zeng et al., 2020) offered deep learning mechanisms not captured in the Doctor AI methodology. The Dipole method uses bidirectional LSTM and attention mechanisms which

could theoretically have better captured the relationships between patient DBCs. The MSAM method leverages two attention encoders to capture both DBC and the lower-level medical code relationships, while being less influenced by irregular visits. Additionally, some of the methods identified in the literature study and Table 5 could have been tailored to the problem at hand and offer different deep learning mechanisms. However, most of the methods only present the important formulas in their work, without disclosing any form of code they have used. Beside the methods of Choi et al. (2016a; 2016c; 2017) only FCNBLA (Wang et al., 2020) and partially MSAM (Zeng et al., 2020) publicly present their Python codes for validation or reuse. In order to compare multiple methods in this research, it would have been necessary to build the algorithm from start to finish relying on the most prominent formulas in the identified papers and make assumptions on the steps taken in the pre-processing of the data. The capabilities of the researcher lack in this regard as the programming skills to implement conceptual deep learning into real-world programmable Python code are not developed enough. Together with the limited time available to conduct this research in the Master Thesis Project format, the ideal situation of comparing different methods was not feasible. The justification of this research, that more validation on machine learning methods needs to be done in order to generalise performance, has been addressed throughout this research. Future research should not only focus on designing new methods to cope with real-world problems, but should also develop the field of machine learning application further in comparing existing methods on different real-world problems in order to find the right solution.

The second limitation can be found in the size and type of selected data for this research. As the registration of DBC data is the financial backbone of any Dutch hospital, the data is stored in a structured and complete manner. However, all of the in the literature study identified and in Table 5 summarized methods relied on patient admissions or visit data. A patient admission or visit in most cases is a collection of more than one diagnosis and treatment and differs in that regard to a DBC which is a recording of the overarching single diagnosis and treatment. As the admission and visiting data is collected by the 21 different hospital specialisms, the way of storing and collecting data is different in every situation. Therefore, the national standardized DBCs offered the most complete, structured and reliable data which was able to represent the overall patient care in VieCuri. The results for this level and type of data in comparison to the Doctor AI (Choi et al., 2016c), Dipole (Ma et al., 2017) and MSAM (Zeng et al., 2020) yielded a dataset which was rich in diversity of labels, but poor in size. In Table 18 the comparison between the size of the four datasets is listed. What becomes clear from Table 18 is that although the total number of recordings in the VieCuri dataset (1.797.467) is bigger than from Dipole (1.055.011) and MSAM (1.301.954), they are stretched over a much longer period of time (7 years in contract to 2 and 1 year(s)). Especially in the number of records per patient per year and number of medical codes per record, the VieCuri dataset significantly lags behind in size when compared to the other three datasets.

Although the one code per recording is the result of working with DBC data instead of admission or visit data, the number of records per patient per year are highly influenced by the number of patient visits. In itself it can be reasoned that the three hospitals of which the other datasets are collected operate for a much larger and densely populated service area than the region where VieCuri operates. However, this does not explain why on average a patient visits around once a year (1,06) in VieCuri, but does so much more frequently in the other three datasets: 6,83 (Doctor AI), 3,57 (Dipole) and 4,45 (MSAM). In having on average only 1,06 records per patient per year, the occurrence of stand-alone events like for example receiving a

flu vaccine in 2012 and treatment for a broken arm in 2016 are much higher. As Choi et al. (2016 p.10) stated in their Doctor AI research: “We have also shown that the patient's visit count and the rarity of medical codes highly influence the performance”. This statement holds as the VieCuri dataset is sparse when compared to the other three researches. Future research could be applied to the impact of data density/sparsity on the performance of machine learning in healthcare research and how to overcome them, as not every hospital is capable of achieving the high density of visits that is used in the three mentioned researches.

Table 18: Dataset comparison VieCuri, Doctor AI, MSAM and Dipole

	VieCuri	**Doctor AI**	**Dipole**	**MSAM**
Total Patients	240.841	263.706	147.810	146.287
Total Records	1.797.467	14.400.985	1.055.011	1.301.954
Records per patient	7,46	54,61	7,14	8,90
Records per patient per year	1,06	6,83	3,57	4,45
Codes per record	1,00	3,22	4,08	5,00
Unique Codes	3.395	38.594	8.522	7.497
Predicted Codes	3.087 till 21	1.183	426	291
Timespan collected data	7 years	8 years	2 years	1 year

Third, it is known that both an aging patient population and patients who consume most current healthcare are the main contributors to healthcare demand and cost (Bates et al., 2014; Callahan & Shah, 2017; Schut et al., 2013). Although these statements are the foundation of this research, external factors which impact healthcare demand are not to be neglected in order to fully mimic patient healthcare demand. As the most prominent impactful example on current healthcare demand, the COVID-19 pandemic has had a major impact on the type and frequency of treatments that are provided by hospitals across the world. For the Netherlands this meant a scale down of regular healthcare treatment and operations to support the growing Intensive Care (IC) demand required by COVID-19 patients (RIVM, 2020). The COVID-19 pandemic, or other examples such as the severe influenza epidemic of the 2017/2018 winter season (RIVM, 2018), or the deathly heatwave of the summer 2019 (CBS, 2019), are all illustrative of external factors that can impact healthcare demand for a single country, region, or hospital significantly. None of the three mentioned external factors could have been predicted or incorporated in the forecasting praxis by only relying on historical patient data.

Following that line of reasoning, the results of the financial impact analysis in paragraph 4.2 show the biggest forecasting fault margins in the domain of Gynaecology, where historical data is not able to predict future birth, as this represents the first patient visit to a hospital. The impact of this stands out as the babies which are born in VieCuri account for more than half of the total specialism cost (53,6%). Although mechanisms as patient birth can be accounted for by using different sources or models, the severity or impact of future external factors as the next pandemic, epidemic, or heatwave are almost impossible to predict. Therefore, the conclusion can be made that even though healthcare demand predictions on historical patient data can be improved onto a more detailed level, the impact of external factors which cannot be extracted from historical data are not to be neglected. Although machine learning could be able to predict the rough numbers of healthcare demand, more research into the impact of external factors is needed to refine these rough numbers to practical use.

5.3 PRACTICAL IMPLICATIONS

For VieCuri Medisch Centrum this research marks the first specialism overarching research into the research field of machine learning and it is thus still in the exploration phase for the hospital. Although optimistic about the potential gains, other Dutch hospitals have also not moved beyond the exploration of machine learning opportunities (ICT&health, 2019). Forecasting healthcare demand is a yearly challenging issue where both hospitals and healthcare insurers are confronted with the struggle of providing healthcare coverage for the whole population while invested in different interests (Schut et al., 2013; Van de Ven & Schut, 2009). Insurers are profit driven while bound to national regulations to decrease the growth of healthcare expenses, whereas hospitals want to provide high quality patient care while keeping expenses within budget. Hospitals which are able to more accurately predict healthcare product consumption one year in advance can achieve more compensation from the healthcare insurers, while not having to treat patients on their own expenses. Being able to more accurately forecast healthcare product demand is a competitive edge in both the yearly negotiation with healthcare insurers as well as the foundation for capacity planning and availability of resources (Kaplan & Porter, 2011).

The results of this research show that performance of machine learning algorithms can vary greatly when methods are applied to new and different datasets and situations. The applied Doctor AI (Choi et al., 2016c) model for predicting the individual healthcare products does not perform well enough to be applied in yearly healthcare demand forecasting praxis for VieCuri. Forecasting future healthcare product demand would generate most insight into the exact product consumption and therefore be a more precise healthcare cost predictor when used for budgeting purposes (Yang et al., 2017; Yang et al., 2018). However, forecasting is not only done within the process of negotiating yearly healthcare insurer budgets and can be executed on different levels throughout the hospital. Therefore, the results on the higher hierarchical level models are still of great value to hospital forecasting praxis.

Dutch hospitals struggle with combining quality and affordability in their healthcare offering (Kroneman et al., 2016; Schut et al., 2013). Being able to forecast the number of patients a hospital specialism will treat for a given year creates the first level of insight in yearly patient flows and shows rough approximation of the involved yearly cost. In showing good performance on both accuracy@3 (75,4%) and recall@3 (72,6%) on the highest hierarchical level model of hospital specialism, this research shows that machine learning can discover patterns in sequential patient data. Although not applied to different use cases in this research, the demand for the specific hospital specialism can provide the first step in more tailored allocation of human and capital resources (Kaplan & Porter, 2011). As doctors, nurses, and treatment facilities are in high demand, being able to forecast on the number of patients can help hospitals better distribute these recourses across the different departments (Bates et al., 2014). Thereby the machine learning model is able to create direct added value while being the starting point for further research into the detailed healthcare demand.

5.4 THEORETICAL IMPLICATIONS

This research contributes to the validation of machine learning methods by applying the Doctor AI methodology (Choi et al., 2016c) to a different dataset in a different setting. Leveraging new machine learning algorithms or innovating on existing methods has contributed to improved performance in a variety of fields of application within machine learning and the healthcare specific domain. However, as acknowledged throughout this research, many of the incorporated research papers indicate that more validation on existing methods needs to be done in order to generalize their performance (Galatzer-Levy et al., 2014; Jiang et al., 2017; Khaldi et al., 2017; Miotto et al., 2016; Prasad & Agarwal, 2014; Roysden & Wright, 2015; Srinivas et al., 2010; Yang et al., 2017; Zlotnik et al., 2015). As the research of Doctor AI (Choi et al., 2016c) claims, their method can be applied from one hospital to the other. The MSAM (Zeng et al., 2020) research validates that by using the Doctor AI methodology as a baseline for their research and achieving similar results in applying the algorithm.

Predicting on medical data has its unique challenges and characteristics, as is discussed in paragraph 2.1. These characteristics include temporality, high diversity and sparseness (Nguyen et al., 2016; Zhou et al., 2017). Combined they confront researchers with problems when data is differs between countries, regions, or hospitals as thereby knowledge cannot be transferred between institutions (Morton et al., 2016). In this research it is shown that although the Doctor AI method can be applied to the dataset of a different hospital (relying on the same ICD10-framework), the nature of the EHR data of VieCuri and the difference in administering medical codes impacts the method's performance significantly. The outcomes of this research show that by decreasing the number of output classes, the model is better able to learn the patient representations, but loses out on detailed information like healthcare product or ICD10-clusters. When observing these results, it should be taken into account that besides the number of nodes per hidden layer, none of the hyperparameter settings of the original Doctor AI algorithm (Choi et al., 2016c) are changed.

This research does not succeed in leveraging the DBC data in order to predict on a healthcare product level. The underlying relationships between visits did not show generalisable results in order to forecast healthcare product demand. The recognition of these underlying relationships thereby remains one of the biggest challenges for deploying machine learning algorithms on EHR data (Barati et al., 2011; Lei, 2017). The Doctor AI setup of this research was not able to overcome the difference in density of patient visits and volume between the different codes in the VieCuri dataset. Both phenomena are a direct effect of working with EHR data, as sequential patient representations are very specific and diverse in nature (Esteva et al., 2019; Solares et al., 2020). Therefore, the customer journey and medical needs for one patient are approached and tailored differently than for another patient. Just as in medical practise the impact of a negative incorrect diagnosis is often more critical than a positive correct one, the field of machine learning application still has hurdles to overcome in order to be applied in mainstream healthcare practice (Miotto et al., 2016; Wiens & Shenoy, 2018). However, the results of this research show that on the highest level of hospital specialism, advancements can develop in small steps on which future improvements can be made.

REFERENCES

Ahamed, F., & Farid, F. (2018). Applying Internet of Things and machine-learning for personalized healthcare: issues and challenges. In 2018 International Conference on Machine Learning and Data Engineering (iCMLDE) (pp. 19-21). IEEE.

Al-Aiad, A., Duwairi, R., & Fraihat, M. (2018). Survey: deep learning concepts and techniques for electronic health record. In 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-5). IEEE.

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's Disease. Frontiers in computational neuroscience, 13, 31.

Bakx, P., O'Donnell, O., & Van Doorslaer, E. (2016). Spending on health care in the Netherlands: not going so Dutch. Fiscal Studies, 37(3-4), 593-625.

Barati, E., Saraee, M. H., Mohammadi, A., Adibi, N., & Ahmadzadeh, M. R. (2011). A survey on utilization of data mining approaches for dermatological (skin) diseases prediction. Journal of Selected Areas in Health Informatics (JSHI), 2(3), 1-11.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123-1131.

Beauvais, B., & Wells, R. (2006). Does money really matter? A review of the literature on the relationships between healthcare organization finances and quality. Hospital Topics, 84(2), 20-29.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). A study of machine learning in healthcare. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 236-241). IEEE.

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ open, 2(4).

Callahan, A., & Shah, N. H. (2017). Machine learning in healthcare. In Key Advances in Clinical Informatics (pp. 279-291). Academic Press.

CBS (2019). More deaths during recent heat wave. https://www.cbs.nl/en-gb/news/2019/32/more-deaths-during-recent-heat-wave

CBS (2020). Statline; Bevolking; geslacht, leeftijd en burgerlijke staat per 1 januari. https://opendata.cbs.nl/statline/?dl=308BE#/CBS/nl/dataset/7461bev/table. 7-9-2020. 16-10-2020.

Che, Z., Kale, D., Li, W., Bahadori, M. T., & Liu, Y. (2015). Deep computational phenotyping. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 507-516).

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.

Chen, Y., Lee, J. Y., Sridhar, S., Mittal, V., McCallister, K., & Singal, A. G. (2020). Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment. Journal of Marketing, 84(3), 1-27.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., ... & Sun, J. (2016a). Multi-layer representation learning for medical concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1495-1504).

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016b). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems (pp. 3504-3512).

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016c). Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference (pp. 301-318).

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 787-795).

Choi, E., Xiao, C., Stewart, W., & Sun, J. (2018). Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In Advances in Neural Information Processing Systems (pp. 4547-4557).

Chung, Y. A., Weng, W. H., Tong, S., & Glass, J. (2019). Towards unsupervised speech-to-text translation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7170-7174). IEEE.

Clark, D. O., Von, M. K., Saunders, K., Baluch, W. M., & Simon, G. E. (1995). A chronic disease score with empirically derived weights. Medical care, 33(8), 783-795.

Coates, A., & Ng, A. Y. (2011). The importance of encoding versus training with sparse coding and vector quantization. In ICML.

Coiera, E., Wang, Y., Magrabi, F., Concha, O. P., Gallego, B., & Runciman, W. (2014). Predicting the cumulative risk of death during hospitalization by modeling weekend, weekday and diurnal mortality risks. BMC health services research, 14(1), 226.

Dahlem, D., Maniloff, D., & Ratti, C. (2015). Predictability bounds of electronic health records. Scientific reports, 5(1), 1-9.

De Jong, A., & Van Duin, C. (2010). Regionale prognose 2009-2040: Vergrijzing en omslag van groei naar krimp. Centraal Bureau voor de Statistiek, Heerlen.

Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A. R., & Hinton, G. (2010). Binary coding of speech spectrograms using a deep auto-encoder. In Eleventh Annual Conference of the International Speech Communication Association.

Dove, H. G., Duncan, I., & Robb, A. (2003). A prediction model for targeting low-cost, high-risk members of managed care organizations. Am J Manag Care, 9(5), 381-9.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68-77.

Du, Z., Yang, Y., Zheng, J., Li, Q., Lin, D., Li, Y., ... & Cai, Y. (2020). Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. JMIR medical informatics, 8(7), e17257.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.

Feng, Y., Min, X., Chen, N., Chen, H., Xie, X., Wang, H., & Chen, T. (2017). Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 770-777). IEEE.

FMS (2018). Hoofdlijnenakkoord medisch-specialistische zorg. Federatie Medisch Specialisten. 25-4-2018. https://www.demedischspecialist.nl/onderwerp/hoofdlijnenakkoord. 30-6-2020.

Finarelli Jr, H. J., & Johnson, T. (2004). Effective demand forecasting in 9 steps: shifts in demand for a hospital's services can occur unexpectedly. Demand forecasting can help you prepare for these shifts and avoid strategic missteps. Healthcare Financial Management, 58(11), 52-58.

Galatzer-Levy, I. R., Karstoft, K. I., Statnikov, A., & Shalev, A. Y. (2014). Quantitative forecasting of PTSD from early trauma responses: A machine learning application. Journal of psychiatric research, 59, 68-76.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., ... & Celi, L. A. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PloS one, 13(2), e0192360.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., & Ranganath, R. (2018). Opportunities in machine learning for healthcare. arXiv preprint arXiv:1806.00388.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

Gupta, V., Sachdeva, S., Bhalla, S. (2020). A Novel Deep Similarity Learning Approach to Electronic Health Records Data. IEEE.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare informatics research, 19(2), 121-129.

Haq, H. U., Ahmad, R., & Hussain, S. U. (2017). Intelligent EHRs: predicting procedure codes from diagnosis codes. arXiv preprint arXiv:1712.00481.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, Z., Dong, W., Ji, L., & Duan, H. (2016). Outcome prediction in clinical treatment processes. Journal of medical systems, 40(1), 8.

Huff, T. J., Ludwig, P. E., & Zuniga, J. M. (2018). The potential for machine learning algorithms to improve and reduce the cost of 3-dimensional printing for surgical planning. Expert review of medical devices, 15(5), 349-356.

IBM Watson for Oncology. (2020). Oncology and Genomics. Available from: https://www.ibm.com/watson/health/oncology-and-genomics/oncology/. 8-6-2020.

ICT&health. (2019). Ziekenhuizen zien potentieel AI, werken aan beleid. https://www.icthealth.nl/nieuws/ziekenhuizen-zien-potentieel-ai-werken-aan-beleid/. 14-2-2019. 8-3-2021.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

Jiang, S., Chin, K. S., Wang, L., Qu, G., & Tsui, K. L. (2017). Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. Expert systems with applications, 82, 216-230.

Jödicke, A. M., Zellweger, U., Tomka, I. T., Neuer, T., Curkovic, I., Roos, M., ... & Egbring, M. (2019). Prediction of health care expenditure increase: how does pharmacotherapy contribute? BMC health services research, 19(1), 953.

Kam, H. J., Sung, J. O., & Park, R. W. (2010). Prediction of daily patient numbers for a regional emergency medical center using time series analysis. Healthcare informatics research, 16(3), 158-165.

Kaplan, R. S., & Porter, M. E. (2011). How to solve the cost crisis in health care. Harv Bus Rev, 89(9), 46-52.

Khaldi, R., El Afia, A., Chiheb, R., & Faizi, R. (2017). Artificial neural network-based approach for blood demand forecasting: Fez transfusion blood center case study. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (pp. 1-6).

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kroneman, M., Boerma, W., van den Berg, M., Groenewegen, P., de Jong, J., & van Ginneken, E. (2016). Netherlands: health system review.

Lapin, M., Hein, M., & Schiele, B. (2015). Top-k multiclass SVM. arXiv preprint arXiv:1511.06683.

Lapin, M., Hein, M., & Schiele, B. (2016). Loss functions for top-k error: Analysis and insights. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1468-1477).

Lei, S. (2017). Predict the Future Hospitalized Patients Number Based on Patient's Temporal and Spatial Fluctuations Using a Hybrid ARIMA and Wavelet Transform Model. Journal of Geographic Information System, 9(4), 456-465.

Li, Y., Qian, B., Zhang, X., & Liu, H. (2020a). Knowledge guided diagnosis prediction via graph spatial-temporal network. In Proceedings of the 2020 SIAM International Conference on Data Mining (pp. 19-27). Society for Industrial and Applied Mathematics.

Li, Z., Wen, L., Liu, J., Jia, Q., Che, C., Shi, C., & Cai, H. (2020b). Fog and Cloud Computing Assisted IoT Model Based Personal Emergency Monitoring and Diseases Prediction Services. Computing and Informatics, 39(1-2), 5-27.

Liang, Z., Liu, J., Ou, A., Zhang, H., Li, Z., & Huang, J. X. (2019). Deep generative learning for automated EHR diagnosis of traditional Chinese medicine. Computer methods and programs in biomedicine, 174, 17-23.

Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. Mis Quarterly, 41(2).

Lin, C., Zhangy, Y., Ivy, J., Capan, M., Arnold, R., Huddleston, J. M., & Chi, M. (2018). Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In 2018 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 219-228). IEEE.

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., & Gao, J. (2017). Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1903-1911).

Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., & Gao, J. (2018). Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 743-752).

Microsoft Project InnerEye. (2020). Assistive AI for Cancer Treatment. Available from: https://www.microsoft.com/en-us/research/project/medical-image-analysis/#. 8-6-2020.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Milovic, B., & Milovic, M. (2012). Prediction and decision making in health care using data mining. Kuwait Chapter of the Arabian Journal of Business and Management Review, 1(12), 126.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6(1), 1-10.

Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., & Kakadiaris, I. A. (2014). A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In 2014 13th International Conference on Machine Learning and Applications (pp. 428-431). IEEE.

Morton, M., Nagpal, S., Sadanandan, R., & Bauhoff, S. (2016). India's largest hospital insurance program faces challenges in using claims data to measure quality. Health Affairs, 35(10), 1792-1799.

Mu, Y., Huang, M., Ye, C., & Wu, Q. (2018). Diagnosis Prediction via Recurrent Neural Networks. International Journal of Machine Learning and Computing, 8(2).

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Nguyen, P., Tran, T., Wickramasinghe, N., & Venkatesh, S. (2016). Deepr: a convolutional net for medical records. IEEE journal of biomedical and health informatics, 21(1), 22-30.

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of global health, 8(2).

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2016). Deepcare: A deep dynamic memory model for predictive medicine. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 30-41). Springer, Cham.

Pietz, K., Ashton, C. M., McDonell, M., & Wray, N. P. (2004). Predicting healthcare costs in a population of veterans' affairs beneficiaries using diagnosis-based risk adjustment and self-reported health status. Medical care, 1027-1035.

Pitocco, C., & Sexton, T. R. (2018). Measuring hospital performance using mortality rates: an alternative to the RAMR. International journal of health policy and management, 7(4), 308.

Prasad, B. R., & Agarwal, S. (2014). Modelling risk prediction of diabetes—A preventive measure. In 2014 9th International Conference on Industrial and Information Systems (ICIIS) (pp. 1-6). IEEE.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

Rahane, W., Dalvi, H., Magar, Y., Kalane, A., & Jondhale, S. (2018). Lung cancer detection using image processing and machine learning healthcare. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-5). IEEE.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Sundberg, P. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 18.

RIVM (2018). Annual report Surveillance of influenza and other respiratory infections in the Netherlands: winter 2017/2018. https://www.rivm.nl/bibliotheek/rapporten/2018-0049.pdf

RIVM (2020). Ontwikkeling COVID-19 in grafieken. RIVM De zorg voor morgen begint vandaag. https://www.rivm.nl/coronavirus-covid-19/grafieken. 16-10-2020.

Rose, S. (2018). Robust machine learning variable importance analyses of medical conditions for health care spending. Health services research, 53(5), 3836-3854.

Roysden, N., & Wright, A. (2015). Predicting health care utilization after behavioural health referral using natural language processing and machine learning. In AMIA Annual Symposium Proceedings (Vol. 2015, p. 2063). American Medical Informatics Association.

Ruan, T., Lei, L., Zhou, Y., Zhai, J., Zhang, L., He, P., & Gao, J. (2019). Representation learning for clinical time series prediction tasks in electronic health records. BMC Medical Informatics and Decision Making, 19(8), 259.

Russel-Puleri, S. (2019). Electronic Health Records GRUs. https://github.com/sparalic/Electronic-Health-Records-GRUs. The City College of New York.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.

Schut, E., Sorbe, S., & Høj, J. (2013). Health care reform and long-term care in the Netherlands.

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In Twelfth annual conference of the international speech communication association.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, 5 (4), 13-22.

Sheets, L., Petroski, G. F., Zhuang, Y., Phinney, M. A., Ge, B., Parker, J. C., & Shyu, C. R. (2017). Combining Contrast Mining with Logistic Regression to Predict Healthcare Utilization in a Managed Care Population. Applied clinical informatics, 8(2), 430.

Shi, J., Fan, X., Wu, J., Chen, J., & Chen, W. (2018). DeepDiagnosis: DNN-based Diagnosis Prediction from Pediatric Big Healthcare Data. In 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD) (pp. 287-292). IEEE.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.

Silverstein, M. D., Qin, H., Mercer, S. Q., Fong, J., & Haydar, Z. (2008). Risk factors for 30-day hospital readmission in patients≥ 65 years of age. In Baylor University Medical Center Proceedings (Vol. 21, No. 4, pp. 363-372). Taylor & Francis.

Smartvision (2019). An overview of the CRISP DM methodology. https://smartvision-me.com/the-crisp-dm-data-mining-methodology/. 21-2-2021.

Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., ... & Conrad, N. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. Journal of Biomedical Informatics, 101, 103337.

Soyiri, I. N., & Reidpath, D. D. (2013). An overview of health forecasting. Environmental health and preventive medicine, 18(1), 1-9.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

STZ Ziekenhuizen (2020). Samenwerkende Topklinische-opleidings Ziekenhuizen. https://www.stz.nl/1087/over-ons/stzziekenhuizen. 20-5-2020.

Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). Readmission rates and their impact on hospital financial performance: a study of Washington hospitals. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 56, 0046958019860386.

Van de Ven, W. P., & Schut, F. T. (2009). Managed competition in the Netherlands: still work-in-progress. Health Economics, 18(3), 253-255.

Van den Berg, M., Heijink, R., Zwakhals, L., Verkleij, H., & Westert, G. (2011). Health care performance in the Netherlands: Easy access, varying quality, rising costs. European Union law and health, 16(4), 27.

Wammes, J. J. G., Tanke, M., Jonkers, W., Westert, G. P., Van der Wees, P., & Jeurissen, P. P. (2017). Characteristics and healthcare utilisation patterns of high-cost beneficiaries in the Netherlands: a cross-sectional claims database study. BMJ open, 7(11).

Wang, W. W., Li, H., Cui, L., Hong, X., & Yan, Z. (2018). Predicting Clinical Visits Using Recurrent Neural Networks and Demographic Information. In 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)) (pp. 353-358). IEEE.

Wang, T., Xuan, P., Liu, Z., & Zhang, T. (2020). Assistant diagnosis with Chinese electronic medical records based on CNN and BiLSTM with phrase-level and word-level attentions. BMC Bioinformatics, 21(1), 1-16.

Wanyan, T., Kang, M., Badgeley, M. A., Johnson, K. W., De Freitas, J. K., Chaudhry, F. F., ... & Wang, F. (2020). Heterogeneous Graph Embeddings of Electronic Health Records Improve Critical Care Disease Predictions. In International Conference on Artificial Intelligence in Medicine (pp. 14-25). Springer, Cham.

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th international conference on evaluation and assessment in software engineering (pp. 1-10).

Xie, Y., Schreier, G., Chang, D. C., Neubauer, S., Redmond, S. J., & Lovell, N. H. (2014). Predicting number of hospitalization days based on health insurance claims data using bagged regression trees. In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 2706-2709). IEEE.

Xie, Y., Schreier, G., Chang, D. C., Neubauer, S., Liu, Y., Redmond, S. J., & Lovell, N. H. (2015). Predicting days in hospital using health insurance claims. IEEE journal of biomedical and health informatics, 19(4), 1224-1233.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 25(10), 1419-1428.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., ... & Garber, J. E. (2017). Using machine learning to parse breast pathology reports. Breast cancer research and treatment, 161(2), 203-211.

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2017). Machine learning approaches for predicting high utilizers in health care. In International Conference on Bioinformatics and Biomedical Engineering (pp. 382-395). Springer, Cham.

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. biomedical engineering online, 17(1), 131.

Yang, S., Zhengy, X., & Yuan, F. (2019). A Patient Outcome Prediction based on Random Forest. In Proceedings of the 4th International Conference on Crowd Science and Engineering (pp. 220-227).

Yoon, J., Davtyan, C., & van der Schaar, M. (2016). Discovery and clinical decision support for personalized healthcare. IEEE journal of biomedical and health informatics, 21(4), 1133-1145.

Yu, K., Lin, Y., & Lafferty, J. (2011). Learning image representations from the pixel level via hierarchical sparse coding. In CVPR 2011 (pp. 1713-1720). IEEE.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Zeng, X., Feng, Y., Moosavinasab, S., Lin, D., Lin, S., & Liu, C. (2020). Multilevel Self-Attention Model and its Use on Medical Risk Prediction. In Pac Symp Biocomput (pp. 115-126).

Zhao, Y., Ash, A. S., Ellis, R. P., Ayanian, J. Z., Pope, G. C., Bowen, B., & Weyuker, L. (2005). Predicting pharmacy costs and other medical costs using diagnoses and drug claims. Medical care, 34-43.

Zhou, C., Jia, Y., Motani, M., & Chew, J. (2017, August). Learning deep representations from heterogeneous patient data for predictive diagnosis. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 115-123).

Zlotnik, A., Gallardo-Antolín, A., Alfaro, M. C., Pérez, M. C. P., & Martínez, J. M. M. (2015). Emergency department visit forecasting and dynamic nursing staff allocation using machine learning techniques with readily available open-source software. CIN: Computers, Informatics, Nursing, 33(8), 368-377.

APPENDIX A – VieCuri Medisch Centrum

VieCuri Medisch Centrum location Venlo was founded in 1983 under the name Sint Maartens Gasthuis from a merger of two smaller local hospitals. Later in 1992 it merged again with the hospital in Venray and from thereon it operates under its current name. VieCuri nowadays is a collective of two hospitals and three regional clinics that all service the North-Limburg region, which inhabits roughly 257.000 people.

The hospital is structured in ten vertical silos of which three provide the core patient care: cluster *Medisch Ondersteunend*, *Snijdend*, cluster *Moeder en Kind* and cluster *Beschouwend*. For the full organisational chart of the hospital see Figure 16 below. Within the three mentioned clusters twenty offered specialties are embedded. As Top Clinical Care (STZ Ziekenhuizen, 2020) labelled hospital VieCuri provides in the three prescribed pillars patient care, education and research. First, the scale of core patient care can be best interpreted by the annual reviews of the last three years which are shown in Table 1. Second, as educational centre it hosts clinical internships as well as a third of the doctors' workforce being in training. Last, as research institute the hospital contributed to the last three business years an average of 185 papers per year in more than 150 different journals.

This Master Thesis is hosted by the Cluster *Financiën*, a collaboration between four departments which are highlighted in Appendix A. The department *Planning & Control* (P&C) is in the lead on business understanding, processes, and budgeting. *Bedrijfsinformatie* (BI) is in control of data gathering and data understanding within the hospital. *Bureau Integrale Capaciteitsmanagement* (BIC) is responsible for the tactical capacity management, advising other departments on business analytics and creating enough capacity for beds and nursing hours within the P&C budgeting. Finally, *Informatiemanagement* (IM) is responsible for the data warehousing and management of information infrastructure. The combination of different fields of expertise and commitment from multiple departments gives confidence in the integration of information from different perspectives.

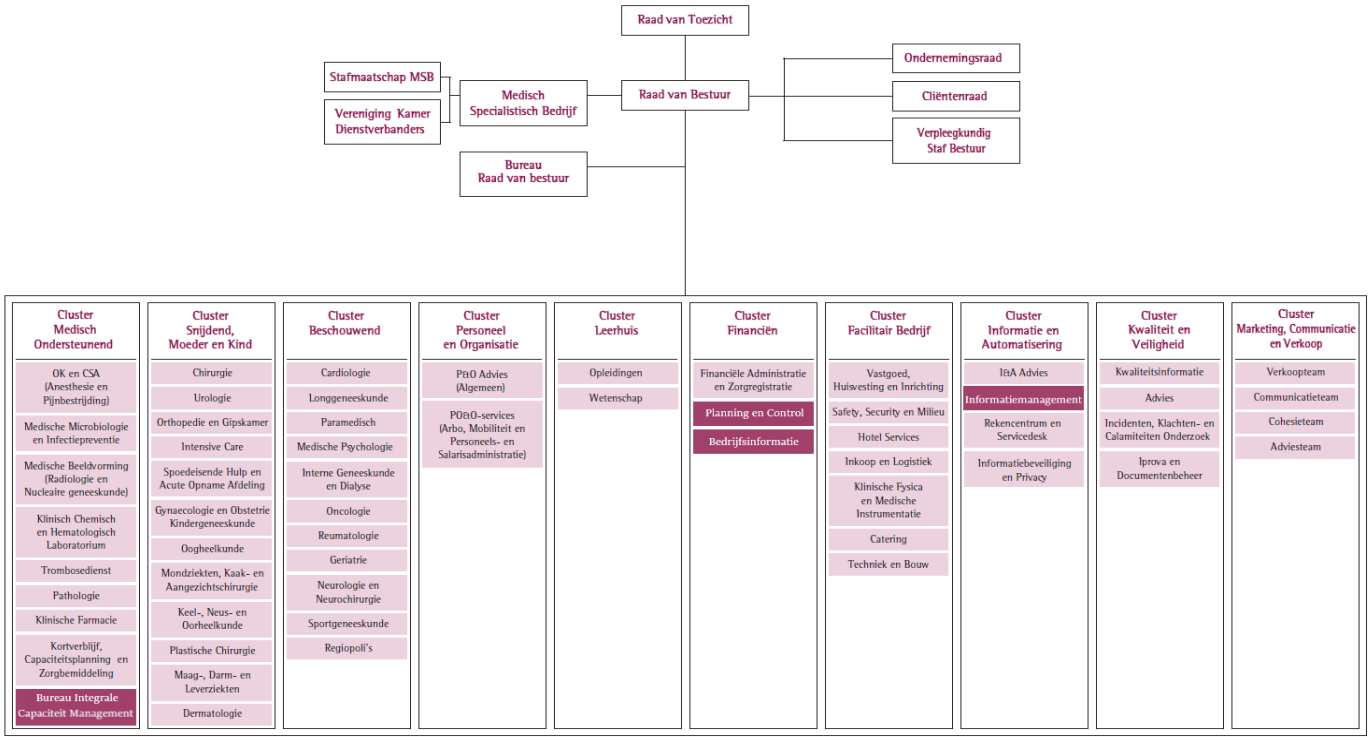

Figure 16: Placement of departments within the organisational chart

APPENDIX B – Literature Review

The literature review which is part of this research has a very focused and narrowed search field as it serves to answer the question which machine learning method is most suitable for handling the problem at hand. The goal of this literature search is the identification of existing methods which have proven effective on a given dataset within the healthcare domain.

B.1 Search Query

The first step of the literature review consisted of the synthesis of search terms in order to effectively narrow and browse the different scientific databases. The search terms which were included can be found in Table 19 and are a result of building on the four recent literature reviews which were used as starting point (Al-Aiad et al., 2018; Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018). Note that by incorporating the "Diagnos*" search term which includes both keywords diagnosis and diagnose as they are used interchangeably. The same holds true for "Record*" for record and the plural records or for the term "Predict*" as therewith all three keywords prediction, predictor, predicting are incorporated. With the selected search terms the field of research, type of data and objected output terms are included. All articles which do not fit one all these three areas are excluded as they do not serve the goal of this review.

Table 19: Database search terms

Search Terms	**Sources**
"Electronic Health Record*" OR "EHR"	Al-Aiad et al., 2018; Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018
"Machine Learning"	Al-Aiad et al., 2018; Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018
"Deep Learning" OR "Neural Networks"	Al-Aiad et al., 2018; Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018
"Diagnos* Predict*" OR "Disease Predict*"	Shickel et al., 2017; Solares et al., 2020; Xiao et al., 2018
"Outcome Predict*"	Al-Aiad et al., 2018; Shickel et al., 2017

The second step is the combination of search terms into a specific search query which is used to browse the databases of Web of Science (WoS) (https://apps.webofknowledge.com), Scopus (https://www.scopus.com) and IEEE/IES Xplore (IEEE) (https://ieeexplore.ieee.org). All three databases were browsed on the combination of title, abstract and keywords. The search query which is displayed below has been used in all three databases. Also, the by the four literature compared methods are included in the set of articles for further investigation. The number of articles which were found in total are shown in Table 20 on the following page.

Search query =

(("Electronic Health Record*" OR "EHR") AND
("Machine Learning" OR "Deep Learning" OR "Neural Networks") AND
("Diagnos* Predict*" OR "Outcome Predict*" OR "Disease Predict*"))

B.2 Selection

The third step in the literature review is the selection of relevant articles which present a method that could be used to better forecast healthcare demand. Therefore, several rounds of assessment of the found articles.

As the research field into the specific subject is relatively young it would have made sense to exclude article on the basis of age, however, none of the databases returned articles older than 2014 (WoS 2017-2020; Scopus 2014-2021; IEEE 2017-2020). Therefore, nothing has been excluded on the basis of publication year. Also, because the research field is so young the emphasis is not only on articles which have been sited at least an x number of times, therefore no articles are dropped due to not enough citations.

The first round of assessment on the results of the three databases is done on the basis of the title and abstract. These give an indication about the topic of the article. If the article does not present a method which predicts any diagnosis/diseases/healthcare outcomes the article is excluded in this round. Also, the article of Gupta et al. (2020) from the IEEE database was excluded, as beside the abstract this was the only non-accessible article. The number of articles which are included after the first round of assessment are shown in Table 20 below.

The second round of assessment is on the remaining articles including the articles which were included from the previously mentioned four literature reviews (LR). The evaluation is done by scanning the articles on the abstract, data descriptive, conclusion, discussion and implications. In this step it is of importance to only incorporate methods which can be generalizable and have therefore shown performance on predicting multiple diagnoses or diseases. Presented methods that only have shown results in form example heart failure, diabetes, bone disease or early mortality are excluded. The number of articles that remain are shown in Table 20.

All unique articles which are left after the second assessment round are displayed in more detail in Table 5 in the main text of Section 2.2.

After the evaluation of the identified methods the discussion which is presented in Section 2.3 follow up to the selection of the two most suitable methods for the problem at hand and are explained in more detail in Section 2.4.

Table 20: Number of articles per literature study phase

Phase	**WoS**	**Scopus**	**IEEE**	**Four LR**
Initial Search Query	33	70	27	-
Unique articles after Initial Search Query	83			-
After First Assessment Round	24	43	12	17
Unique articles after First Assessment Round	58			
After Second Assessment Round	4	10	1	6
Unique articles after Second Assessment Round	13			
After Final Assessment	1	2	0	2
After discussion of methods for the problem at hand	3			

APPENDIX C – Performance Visualisation of analysed models

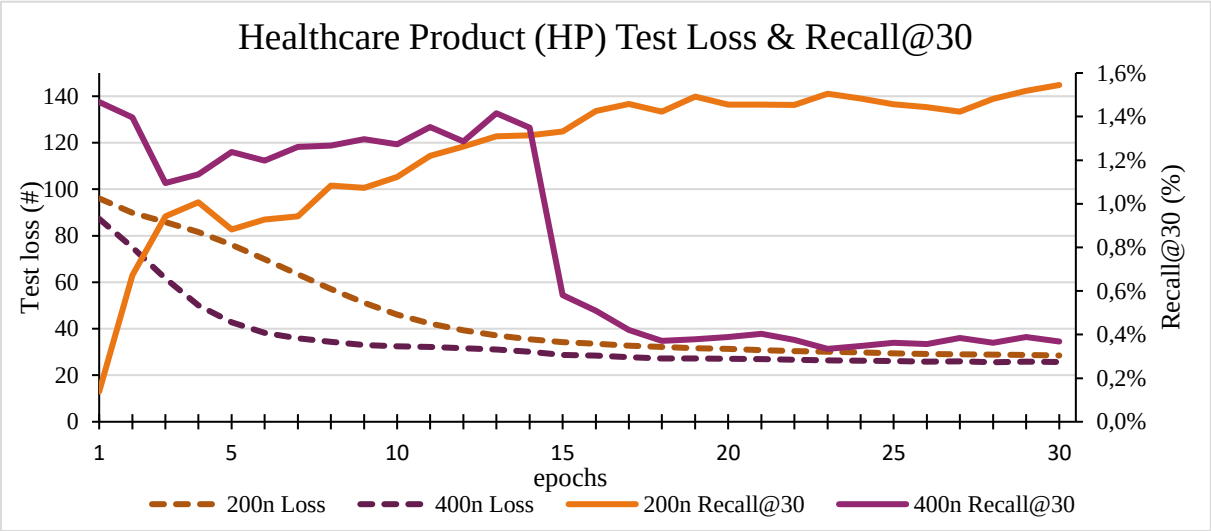

Figure 17: Loss and recall@30 for healthcare product (HP) model (n=3.087)

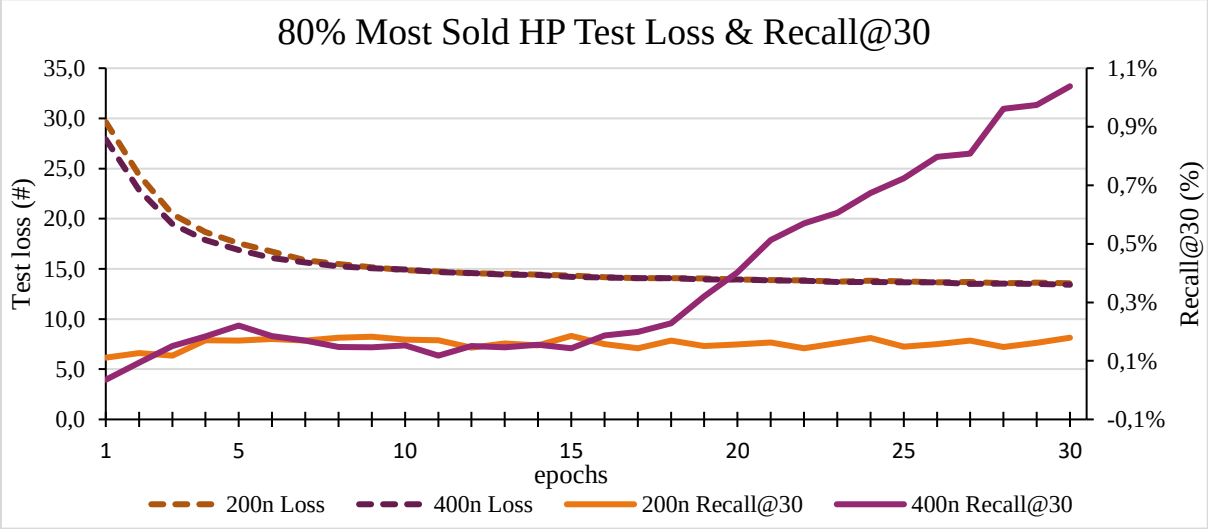

Figure 18: Loss and recall@30 for 80% most sold HP model (n=1.027)

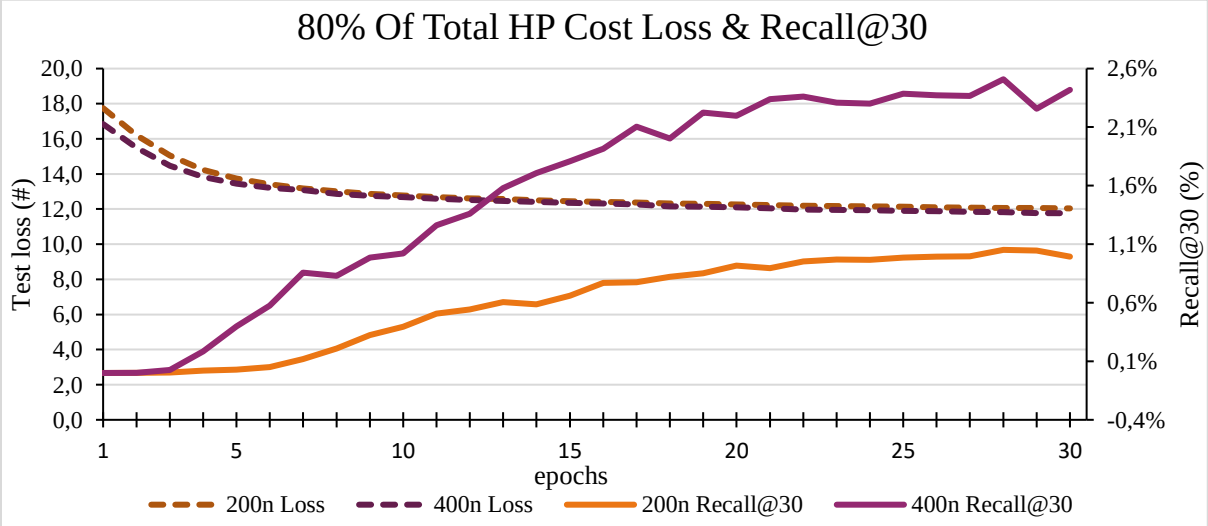

Figure 19: Loss and recall@30 for 80% of total HP cost model (n=519)

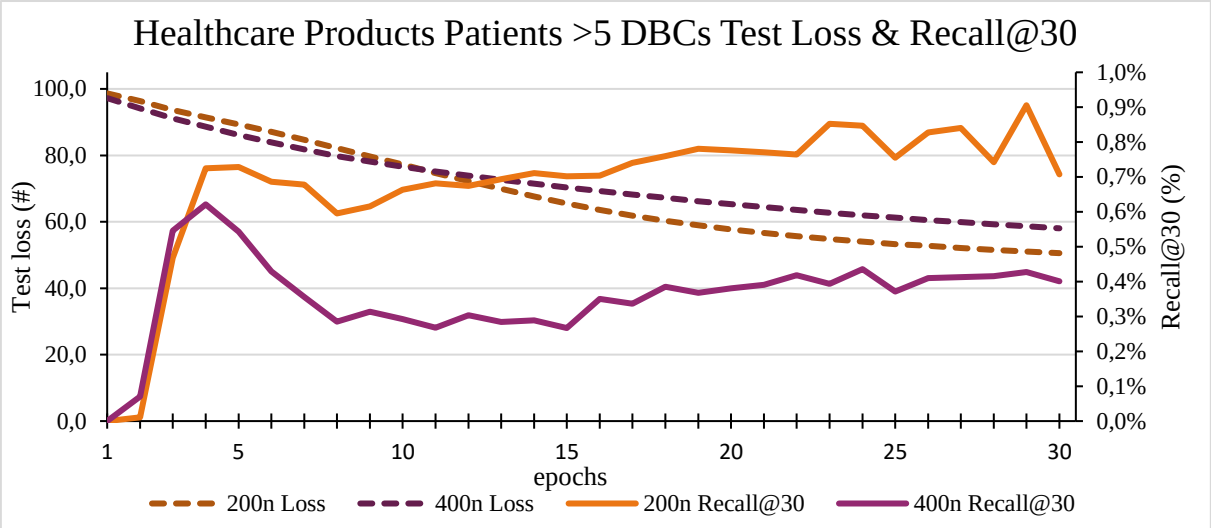

Figure 20: Loss and recall@30 for patients >5 DBCs model (n=3.081)

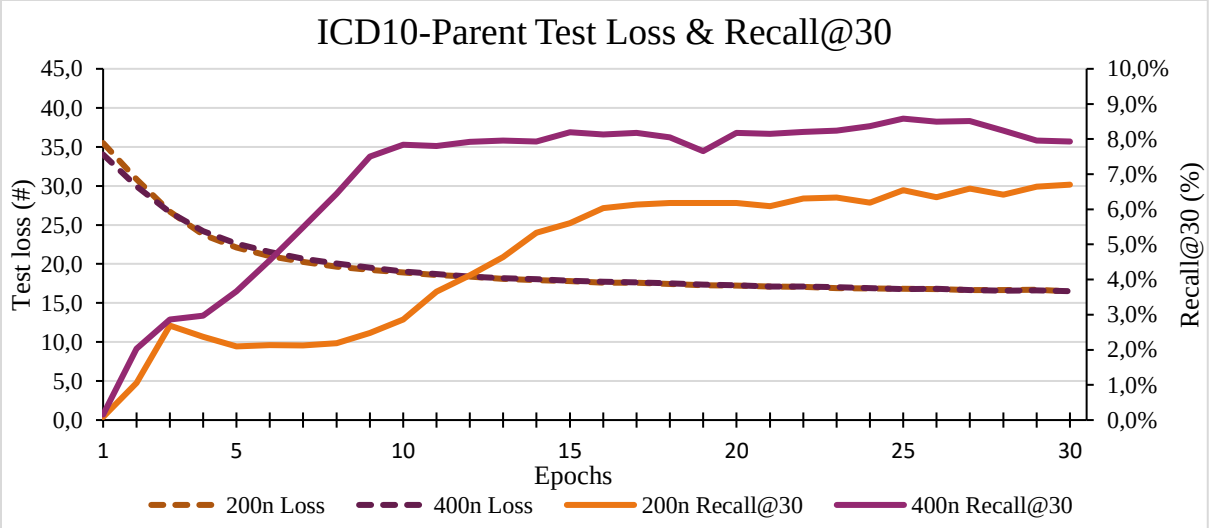

Figure 21: Loss and recall@30 for ICD10-parent model (n=1.132)

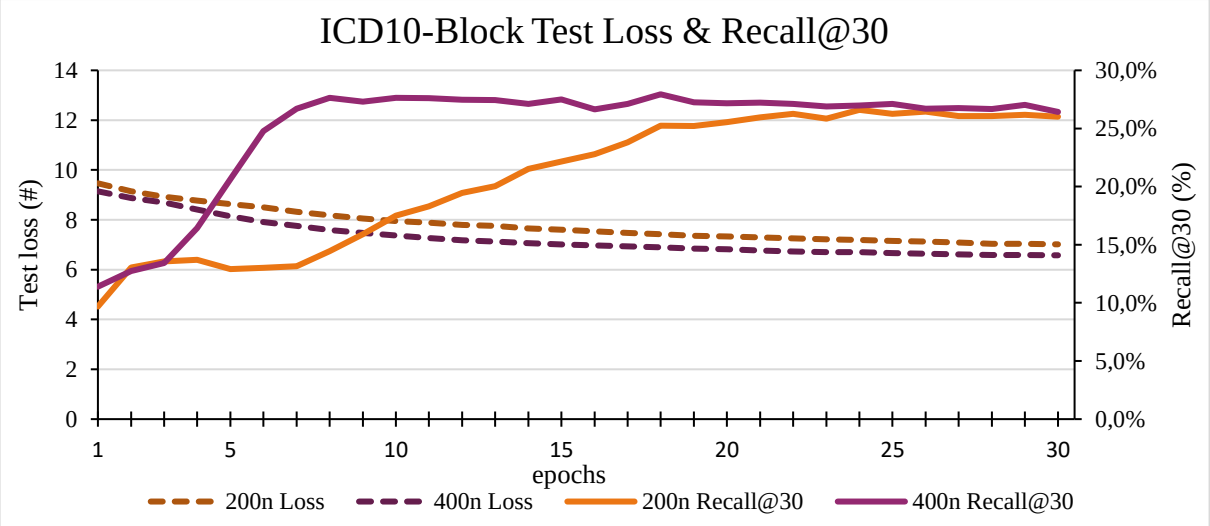

Figure 22: Loss and recall@30 for ICD10-block model (n=199)