Knowledge Discovery in Clinical Psychiatry

Vincent Menger

Cover design: evelienjagtman.com

This work was partially funded by the ZonMw program Kwaliteit van Zorg: Actieonderzoek Innovatieve Zorg.

Knowledge Discovery in Clinical Psychiatry
ISBN: 978–90–393–7170–1

Knowledge Discovery in Clinical Psychiatry

Learning from Electronic Health Records

Kennisontdekking in de Klinische Psychiatrie

Leren van Elektronische Patiëntendossiers
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 2 oktober 2019 des middags te 2.30 uur

door

Vincent Jorn Menger

geboren op 16 juli 1989
te Franeker

Promotoren:	Prof. dr. S. Brinkkemper
	Prof. dr. F.E. Scheepers
Copromotor:	Dr. M.R. Spruit

Acknowledgments

Four years ago, when I was presented with the opportunity to pursue a PhD combining computer science and psychiatry, I did not hesitate for long. At the moment, I was sure that I wanted to use my technical background to make a contribution in a societally relevant domain—and in that regard one cannot find many fields more suitable than psychiatry. I have learned a lot in the years that followed, both personally and professionally, with this dissertation as the result of my scientific endeavors. As such, I am very proud to present it here.

What now remains is to express my sincere gratitude to some of the inspiring people I met during the years, and without whom this work would have never come quite close to completion.

First and foremost, I should thank my supervisors. Sjaak, thanks for guiding this research and making your vast experience and knowledge available to me. Knowing I could turn to you for advice has often been an assuring thought. Floor, thanks for your confidence in me. This dissertation would not have existed without your pioneering ideas and dedication, and I cannot recall one meeting we had that did not increase my motivation to continue. Marco, thanks for teaching me many valuable academic lessons, for allowing me all freedom in choosing my research directions, and for your patience with me at the start of this project. I hope that some of your tenacity and determination have also reflected on me during the past years.

Many of my UMCU colleagues have contributed to this work in one way or another. Karin, thanks for your endless support and the great collaboration. Your efforts in the bigger project opened many doors, greatly boosting my progress as well. Femke, thanks for always being a helpful and optimistic person. Zimbo, thanks for your willingness to help with all my whimsical data needs. Jonathan, thanks for coming by at the right moment and advancing our project. Kees and Saskia, thanks for being enthusiastic fans of my work for the last year. Roel and Eline, thanks for the seamless and pleasant collaboration. Elisabeth, Egge, Jeaphianne, Boris, Roberto, Tamar, and others who came and went during the years, thanks for all the discussions, ideas and laughs we shared. Finally, I am especially grateful to all practitioners who have made some of their scarce time available to our experiments.

Even more fortunately, a short bicycle trip across de Uithof away, there were also always my UU colleagues to count on. Erik and Garm, thanks for keeping my spirits up when I just started. Wienand, thanks for always being available to share our achievements and frustrations. Thanks to all fellow PhD candidates, Armel, Bilge, Chaïm, Ian, Ingy, Noha, Michiel, Raj, and Shaheen, and all other colleagues, Fabiano, Jan-Martijn, Matthieu, Marcela, Sergio, Sietse, Slinger, and Veronica, for all the encouraging chats and valuable contributions.

To all students who participated in a research project, especially Liset, Chaïm, Wouter and Nikki, thanks for providing your own unique outlook on things that have inspired many ideas in this dissertation.

I also owe a lot of gratitude to all of you who took my mind of things outside of work—too many to name individually—by sharing dinner, cycling, running, climbing trees, making music, discussing my research, not discussing my research, or simply by drinking a beer. Hopefully, I will now again be able to make up for some of the time I spent behind my laptop.

Ines, thanks for your never-ending patience with my busy work life, for always showing me the bright side of things, and for all the awesome adventures we have shared so far. I am already looking forward to the ones that are still to follow. And finally, Ben, Wilma, Malin, and Jesler, thanks for raising me to the person I am today and for your unconditional support. For that I am truly grateful.

— Vincent, August 2019

Contents

1 Introduction

Modern day health care, predominantly based on medical science, has been able to achieve remarkable progress in improving people's health and well-being (Kuper and D'Eon, 2010). Medical research has created a vast body of knowledge over the past century, drastically improving our understanding of diseases. On a daily basis, health care organizations bring together patients with physical or mental impairments, and health care professionals of various disciplines who draw on this body of knowledge. By combining knowledge, clinical expertise, and the patient's symptoms and experience, they are enabled to establish a diagnosis and select an appropriate treatment (Wyatt, 1991).

Despite the general progress that has been made, further improvement in the quality of care is imperative (Chassin and Galvin, 1998). Not every field within medicine has developed the same improvement in their patients' health—in many medical specialties, there still exist significant gaps in the knowledge and understanding of disease, and insights in how to treat them are often far from complete (Krumholz, 2014). Additionally, the cost of health care has steadily risen in many countries worldwide, where it is soon expected to reach unsustainable levels (Appleby, 2012). For this reason health care is under pressure to simultaneously increase quality of care and decrease its costs (Cohen and Siegel, 2005). To unify these two conflicting objectives, innovative approaches are needed (Murdoch and Detsky, 2013; Obermeyer and Lee, 2017). One potential source of innovation in health care that has emerged in recent years, is finding and using information that is hidden in repositories of Electronic Health Records (EHRs), in which practitioners have registered information about treatment of individual patients (Coorevits et al., 2013; Jensen et al., 2012; Raghupathi and Raghupathi, 2014).

Health care organizations have increasingly adopted EHRs in the past decade, which contain health information about patients in a digital format, accessible and editable by relevant caregivers within the organization. Several initiatives, such as the 2009 Health Information Technology for Economic and Clinical Health act in the USA (HITECH, 2009), have caused a strong rise in EHR adoption. Both in primary care and in hospital settings, a majority of Western health care organizations now uses an EHR (Adler-Milstein et al., 2015a; Rittenhouse et al., 2017). In the USA alone, EHRs are used to digitally

document at least one billion patient visits every year (Hripcsak and Albers, 2013), and even developing countries are increasingly focusing on digitization of their health data (Bram et al., 2015). The primary reason for employing an EHR is documenting treatment, both for enhancing patient care and for reimbursement (King et al., 2013). Which data is present in a patient's EHR depends on individual ailments, but almost always include elements such as diagnosis, medication prescriptions, lab values, medical imaging, genetics, billing codes, unstructured clinical notes, and many more.

Inside these EHRs, implicit information may be hidden on an aggregated level, in addition to the facts of treatment of an individual patient (Bates et al., 2014; Lee and Yoon, 2017). This knowledge can be made explicit trough a knowledge discovery process, which Frawley et al. (1992) defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". In such a process, state of the art techniques from various computer science domains, such as statistics, machine learning, and data visualization are applied to integrate and analyze the various data in large datasets of EHRs. Using knowledge discovery techniques has gained traction in several domains over the past years, including the biomedical domain (Yoo et al., 2011; Zhu and Zheng, 2018). Applying knowledge discovery techniques to EHRs has often been anticipated as a potential source of innovation, and over the past years an increasing number of studies based on large scale EHR data have been reported (Milovic, 2012; Yadav et al., 2018). For the time being however, there are many remaining open questions and barriers regarding the usage of EHR data to improve care, both concerning the contents and quality of data, and concerning the technical, organizational, and ethical aspects of utilizing such datasets (Lokhandwala and Rush, 2016; Priyanka and Kulennavar, 2014). For this reason, EHRs are still underused for research purposes (Obermeyer and Emanuel, 2016). Such challenges need to be addressed before knowledge discovery in EHRs—the focus of this dissertation—can improve care on a large scale.

1.1 Analyzing Electronic Health Records

Today's most common framework to practice medicine is called evidence based medicine, which Sackett et al. (1996) defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients". Within the evidence based medicine framework, various methods for generating evidence are currently employed, the strength of which is typically ranked in various levels. The commonly referenced Oxford CEBM

Levels of Evidence defines five of such levels (OCEBM Levels of Evidence Working Group, 2011). The lowest levels five and four comprise expert opinion and case series. The next levels three and two describe two observational study designs, case-control studies and cohort studies respectively. In both types of studies cases and controls are incorporated, but case-control studies focus on outcome and are retrospective, while cohort studies are designed from the perspective of an exposure and are often of a prospective nature. The first and highest level describes Randomized Controlled Trials (RCTs), the current gold standard for obtaining evidence. A RCT uses an experimental study design in which participants are randomly assigned to a group that is subject to an intervention, and a group that is not. In the majority of studies, *study participants* are enrolled and subject to various measurements, in order to obtain relevant data that is then further analyzed (Institute of Medicine, 2010).

Reusing EHRs as a research repository can offer various benefits over traditional methods of data collection, aside from the already mentioned argument of its increasing availability. Below, we will expand to three such advantages, related to sample size, selection bias, and the possibility of exploration.

Firstly, reusing EHRs enables easier creation of research datasets with adequate *sample size* compared to traditional studies, both in number of patients and in number of measurements (Kohane, 2011; Ramakrishnan et al., 2010). Both in observational and experimental studies, initial participant recruitment, and participant availability at possible follow ups are major bottlenecks for obtaining reliable results (Probstfield and Frye, 2011). Analyzing EHRs can help overcome this problem, since even small health care centers with late EHR adoption currently store data of various types on a number of patients that quickly runs into the thousands, surpassing the size of a typical RCT. Because treatment often lasts over a longer period of time, the majority of this data is longitudinal by default.

Secondly, EHR data is less affected by *selection bias*, a problem that occurs when the population of a study does not match the actual clinical population (Hernán et al., 2004; Pannucci and Wilkins, 2010). For traditional studies, in- and exclusion criteria specify whether participants with specific demographics, diagnoses, comorbidities, and additional relevant factors are eligible for inclusion. Because studies often aim to investigate some phenomenon in its simplest or purest form, they often lack participants that deviate from the 'typical patient', which decreases applicability of results in clinical practice (Lee et al., 2007). Although healthy subjects are typically not present in EHRs, they contain data about all patients that were treated, including those that would ordinarily not meet study inclusion criteria.

Finally, in contrast with the high cost and time associated with traditional data collection, the availability of EHR data allows more straightforward *exploration* in datasets (Angus, 2015; Yadav et al., 2018). This can ultimately lead to uncovering of novel and unexpected patterns and associations that were not necessarily hypothesized in advance by researchers (Khoury and Ioannidis, 2014). Here, EHRs and RCTs can eventually be combined, by using the EHR data to generate hypotheses, and subsequently testing the hypotheses that show most promise or are supported by multiple EHR datasets in RCTs.

Naturally, using EHRs for analysis is not without its challenges. Measurements in EHRs itself are multimodal, irregularly sampled, contain missing values, and can contain significant biases (Hripcsak et al., 2011). Additionally, there are no research protocols governing their accuracy (Zakim and Schwab, 2015). There are technical challenges, such as identifying the best way to preprocess, store, and analyze such data (Hersh et al., 2013b). Furthermore, analyzing EHRs is performed within a health care organization, where additional organizational challenges exist. This for example includes changing the mind set of medical professionals to embrace novel types of research, and collaboration between practitioners and data experts (Ellaway et al., 2014). Ethical considerations, finally, such as patient privacy and legal issues, require serious attention before EHRs can successfully be analyzed (Chawla and Davis, 2013). Addressing such challenges is no straightforward task, however, when these challenges can eventually be addressed, knowledge discovery in EHRs holds exciting opportunities for research in medicine.

Under the conviction that medical research should ultimately aim to improve care for the individual patient, the goal of this research is twofold. Firstly, we aim to identify and resolve challenges associated with using EHRs for data analysis. Secondly, we aim to apply state of the art knowledge discovery techniques to EHRs in order to obtain new knowledge, that can ultimately lead to improvements in care. The individual chapters of this dissertation all intend to contribute to this goal. In the next section of this introduction, we will first introduce the medical research domain, where we will scope our research to the field of psychiatry. We will then elaborate on some of the computing research domains, of which methods and techniques are used and contributed to. Finally, we introduce the main research question and related research questions, the main research methods that were used to investigate them, and outline the rest of the chapters of this dissertation.

1.2 Medical Research Domain of Psychiatry

Within the medical research domain in which this dissertation will operate, we focus on the medical specialty of psychiatry. Psychiatry is the field of medicine that specializes in mental health care, that is, the diagnosis, treatment, and prevention of mental disorders. While mental disorders may partially be the result or cause of physical issues, psychiatric symptoms primarily revolve around affect, behavior, and cognition rather than physical complaints. We will describe three challenges that both research and treatment in psychiatry face, making it a good case for investigating how analyzing EHR data can contribute to improving care (Torous and Baker, 2016; Passos et al., 2016). These three challenges are related to the biomedical model, psychiatric diagnosis, and the problem of objective measurement.

Firstly, the predominant paradigm for research and treatment in psychiatry, the *biomedical model*, has been criticized as too narrow, and insufficiently able to provide translational value (Deacon, 2013). While mental disorders have been studied for centuries, since the 1960s psychiatry has predominantly been viewed in light of the biomedical model (Double, 2002). It assumes that psychopathology is a direct result of a person's biology, most notably at the gene, cell, and neural circuit level. The biomedical model partially came to prominence due to simultaneous advancements in psychopharmacology, which coincidently discovered several new drugs that were able to alter a person's mental state. Prescribing medication can often be considered an important part of treatment within the biomedical view, for example explained by restoring chemical imbalances in the brain (Andreasen, 1985). A notable criticism of the biomedical model is that it leaves little room for social and psychological dimensions of disease, yet concurrent models that do consider these factors are less widespread (Beckett, 2017). Many research efforts within the biomedical model have attempted to explain the aetiology and course of mental disorders by studying brain structure and function using EEGs and MRIs, and by studying genetics using DNA sequencing techniques, with varying results (Deacon and Lickel, 2009). While this research has undoubtedly taught a lot about brain structure, genetics, and their relations to mental disorders, critics point out that this scientific knowledge has so far demonstrated insufficient potential to inspire new forms of treatment. Especially in recent years, a growing number of critics argue that current biomedical research in psychiatry lacks results that ultimately benefit patients (Dean, 2017).

Secondly, *psychiatric diagnosis*, typically performed based on the Diagnostic and Statistical Manual of Mental Disorders (DSM), is often criticized for

its lack of reliability and validity (Adam, 2013; Cuthbert and Insel, 2013). The DSM consists of a set of disorders, along with strict guidelines indicating under which conditions a diagnosis is applicable, determined by deliberation among experts. The first version, published in 1952, intended to provide a categorization of mental disorders to standardize the diverse usage of terminology among practitioners, by describing roughly 60 disorders (American Psychiatric Association, 1952). Over the years, new disorders were added, and criteria were broadened and narrowed, resulting in the current and fifth iteration which describes over 300 disorders (American Psychiatric Association, 2013). Although the DSM is the current standard for diagnosis, is not without controversy. Inter- and intra-rater reliability for performing diagnosis for instance is low for several diagnoses (Brown et al., 2001; Silverman et al., 2001). Validity of the DSM is often debated as well, since it is known that patients with the same diagnosis can exhibit very different symptoms, and many symptoms cut across several disorders (Hilsenroth et al., 2000).

Finally, where *objective measurements* of patients' characteristics are often able to guide the clinical pathway from symptoms to diagnosis in general medicine, such biomarkers are largely absent in psychiatry (Venkatasubramanian and Keshavan, 2016). Objectively measurable characteristics, such as lab values or vitals, have hardly been linked to psychiatric phenotypes (Cristea et al., 2019). Of characteristics that can be relevant in clinical practice, such as symptoms, behavior, and outcome, objective measurement is often problematic (Insel, 2017). Item checklists, consisting of multiple items that are scored by caregivers or patients, are one attempt to record such characteristics in a structured way. However, the nature of these instruments limits their ability to capture the complex dynamics of real patients (Möller, 2009). A large part of relevant information is captured in unstructured formats, for instance in disease history, patient biographies, descriptions of life events, and family anamnesis. These more narrative types of data have almost exclusively been used for research on a small scale after annotation or extraction of information by humans (Abbe et al., 2015). In the end, the difficulty of accurate measurement has resulted in psychiatrists often being left practically without diagnostic or prognostic tools for predicting optimal treatment or outcome (Lakhan et al., 2010).

Today, the percentage of the population that could benefit from mental health care far exceeds the current treatment capacity of psychiatry (Collins et al., 2011). Additionally, psychiatry especially has difficulties treating patients with complex psychopathology, comorbidities, or chronic mental issues (Kathol et al., 2009). Where the biomedical model of psychiatry and the DSM classification of disorders—widespread in research—have insufficiently been

able to inspire innovation in psychiatry, the call for new research approaches has increased. The new field of precision psychiatry for instance aims to "approach treatment and prevention taking into account each person's variability in genes, environment, and lifestyle" (Fernandes et al., 2017). In this context, knowledge discovery in EHRs is well suited to contribute to innovate in psychiatry, both in obtaining new knowledge about mental disorders and their treatment, and in supporting psychiatrists, nurses, and therapists in applying this knowledge to individual patients.

Throughout this dissertation, we will often use data that is routinely registered in the EHRs of patients that were treated at the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in the Netherlands. This department offers both secondary care, for patients directly referred by primary care such as general practitioners, and tertiary care, for patients who need highly specialized care. There are six separate wards, five of which are closed, with a total of roughly 60 inpatient treatment positions. Combined with outpatient treatment, annually roughly 2,000 unique patients are treated. The department consists of four care pathways, *zorglijnen* in Dutch, each with their own focus on a certain patient population. The first care pathway treats patients with the highest mean age, and predominantly focuses on depressive and anxiety disorders. The second care pathway has two separate wards for adolescents and for adults, and focuses on acute care, not limited to specific disorders. The third care pathway focuses on adolescents and young adults, and treats most psychotic disorders and personality disorders. The fourth and final care pathway focuses on developmental disorders, mainly treating children and adolescents, and includes an open and a closed ward. The Department of Psychiatry has employed a patient-centric EHR since 2011, in which all information relevant for treatment is registered by caregivers. Although the department does not treat every possible disorder equally often, the patient population it sees is quite diverse, making it a good case for investigating how analyzing EHR data can contribute to improve psychiatric care.

1.3 Computing Research Domains

In addition to the medical research domain of psychiatry, this dissertation employs and contributes to methods and techniques from various research domains within computing sciences. Most notably, this includes knowledge discovery, machine learning, and natural language processing, which will be introduced in the next sections to benefit the understanding of this dissertation.

1.3.1 Knowledge Discovery

The type of research of this dissertation can best be characterized as knowledge discovery, a multidisciplinary research field that focuses on discovery of interesting insights based on data. Frawley et al. (1992) defined knowledge discovery as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". The term was further popularized by Fayyad et al. (1996), who introduced the Knowledge Discovery in Databases (KDD) process, consisting of five steps from data to knowledge. The attention knowledge discovery has received is strongly connected to the increasing availability of data and computing power since the 1990s, inspiring development of computational methods and theories to extract knowledge from these datasets. The analysis step in the KDD process is called data mining, the process of finding patterns in large datasets. Knowledge discovery approaches have been applied in a wide range of domains, such as retail, transport, communication, engineering, and medicine. In this research, we do not necessarily limit ourselves to discovering knowledge in the form of explicit facts, but also investigate other types of useful insights that can improve care. This also includes predictive modeling, where retrospective datasets are used to model the relation between a number of informative patient characteristics and future outcomes. Such trained models can subsequently be used prognostically, supporting decisions that clinicians need to make during the care process.

In order to structure a knowledge discovery project, several supporting process models exist. In this research, we will use the Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). CRISP-DM was chosen over other process models, such as KDD, and the Sample Explore Modify Model and Assess (SEMMA) model (Azevedo and Santos, 2008), because of its strong organizational component. This makes it well suited for the case of analyzing EHRs within a health care organization. The CRISP-DM is the most widely adopted knowledge discovery model in industry (Onwubolu, 2009).

CRISP-DM consists of six distinct phases (Figure 1.1). First, in the *domain understanding* phase, the domain in which a specific analysis project is situated is explored, in order to gain understanding of the organization in which the project is embedded. The original work called this the business understanding phase, but it has later been referred to as domain understanding as well, which suits our project better. In the subsequent *data understanding* phase, data relevant to the project is accessed, and exploratory analysis is performed to understand its elements. In the *data preparation* phase, preprocessing and transformation operations are applied to the data, in order to make it suitable

for the subsequent *modeling* phase. In this phase, one or more statistical or machine learning models are applied to the prepared dataset. The results of the modeling phase, such as new knowledge or insights, are evaluated in the *evaluation* phase, usually in collaboration with domain experts. In the final *deployment* phase, potential results that are positively evaluated can be introduced into practice, where end-users can make use of these results.

When conducting data analysis, it is often necessary to change back and forth between the different phases, based on insights that were obtained in a later phase. It is furthermore important to note that the CRISP-DM has an iterative character, using obtained results as input for potential subsequent iterations.

The steps of the CRISP-DM are often used throughout this dissertation. To answer the research questions of this dissertation, to be introduced in Section 1.4, our approach is often structured along the CRISP-DM, both implicitly and explicitly. Additionally, these research questions are mapped onto the phases of CRISP-DM in Figure 1.1, where the contribution of each research question in the context of the entire dissertation can be seen.

1.3.2 Machine Learning

Strongly related to the concept of knowledge discovery is the concept of machine learning. It comprises a set of statistical techniques, that enable computers to carry out tasks based on patterns in data, rather than being explicitly programmed to do so. Computers being able to learn from data triggers the imagination of many. To outsiders, the field may seem to be somewhat shrouded in mystery, possibly owing to the fact that machine learning is often mentioned in both scientific literature and traditional media as a solution to many societal problems. However, machine learning is a discipline that is solidly founded in mathematics and statistics, dates back several decades, and has a very active research community. It has already contributed in many areas, and has the potential to contribute in many more.

The idea of machine learning originated from the field of artificial intelligence, which studies how computers can perform human tasks. Among the first machine learning approaches were computer programs that learned to play checkers, based on books with annotated checkers moves (Samuel, 1959). An important milestone in machine learning was the invention of the perceptron, a learning algorithm that produces an output based on a linear combination of its input features (Rosenblatt, 1958). In later years the limitations of the perceptron overshadowed its initial enthusiastic reception, and artificial intelligence research was mostly devoted to learning based on symbolic systems. In

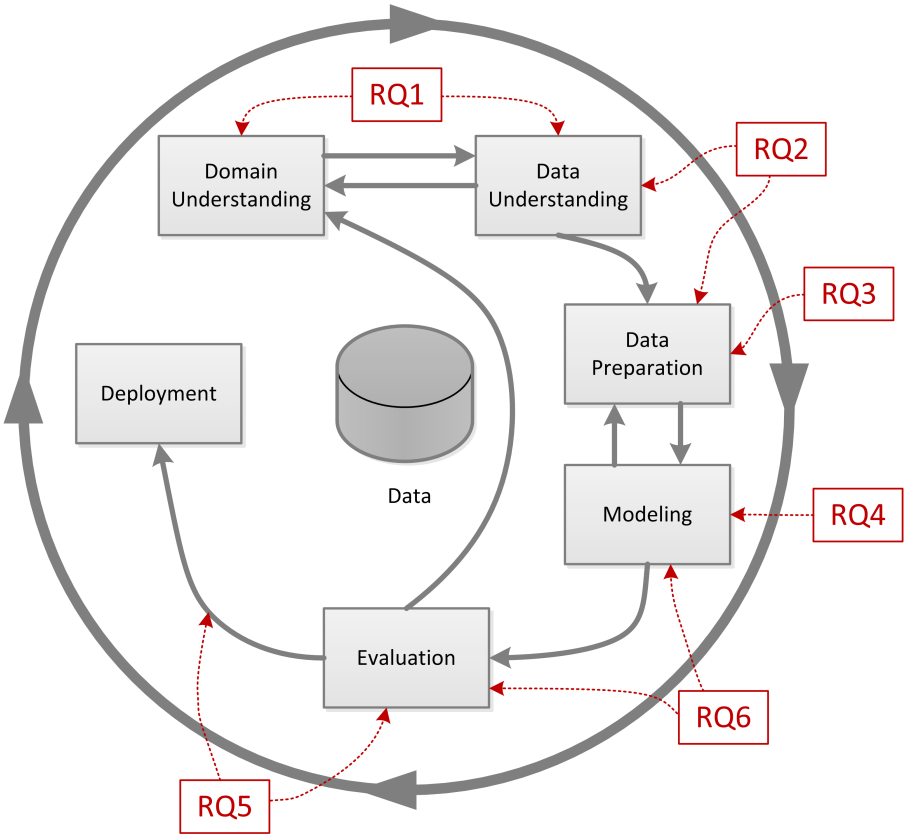


Figure 1.1: The six phases of the CRISP-DM process model, along with a mapping of the six Research Questions of this dissertation onto these phases. Adapted from Chapman et al. (2000). Abbreviations: RQ = Research Question.

the 1980s developments in artificial neural networks, more complex variants of the perceptron, rekindled the interest in machine learning (White, 1989). The invention of various new algorithms together with the increasing availability of data subsequently enabled the use of machine learning in various fields, such as computer vision, information retrieval, speech recognition, and bioinformatics. Deep learning models, essentially increasingly complex variants of neural networks, have lately been able to further improve state-of-the-art results of various problems (LeCun et al., 2015). The ideas of deep learning were already conceived in the 1990s, but availability of computational power and increasingly large datasets have made applying them feasible only very recently.

The most common distinction in machine learning is that between *supervised* and *unsupervised* learning. In the first case, each data point has a known outcome that can be regarded as a ground truth. The task of the learning algorithm is to learn a representation that is able to infer the correct output for each example, often in the form of a label or a real valued number. Unsupervised learning does not make use of a ground truth, but is tasked with learning structure or patterns that are potentially hidden in a dataset. Clustering algorithms for example aim to find latent subgroups of data points that are similar to each other, and dissimilar to other data points.

Although understanding the principles of most machine learning algorithms requires only a basic understanding of mathematics, successfully applying them to real world datasets is much more complex (Salzberg, 1997). What makes this endeavor challenging is for instance evaluating the performance of a model. To use a single dataset for both learning and evaluating a model, a strict distinction between data used for training and testing needs to be made. For this purpose techniques like splitting the dataset or cross validation procedures are used, but the bias-variance dilemma, data leakage issues, and other pitfalls lurk behind every corner (Forman and Scholz, 2010). When a machine learning model is eventually ready for implementation, additional challenges such as potential model bias and model explainability appear. These challenges need ample consideration before a machine learning project can be described as successful.

In this work, we apply supervised machine learning techniques in Chapters 5 and 6, when we look at adopting a data-driven strategy to assess violence risk of patients. In Chapter 7 we use unsupervised machine learning techniques, in the form of cluster ensembles, to identify patient subgroups with interesting characteristics relevant for treatment.

1.3.3 Natural Language Processing

An important part of EHRs consists of unstructured data, most notably in the form of free text in this research. The tradition of registering information in free text goes back to the time when health records were not electronic, and handwritten notes were the physician's method of choice to document their patients well-being. There are two main reasons for this preference to persist in the digital age (Rosenbloom et al., 2011). First, using free text allows physicians to write along the lines of their thought process, which is often done in a semi-structured way. Secondly, unstructured data are able to describe patients' characteristics and well-being with more nuance, and easily allows including additional custom information that cannot be captured in a structured form. On the one hand, data that is captured in an unstructured format can be considered a nuisance, for it is not always straightforward to apply data analysis techniques to this type of information. On the other hand, when the appropriate techniques are applied to them, unstructured data may well be able to provide insights that were never attainable when relying on structured data alone.

Natural Language Processing (NLP) techniques are designed to automatically process human language data—written text in the case of this research. Over the past decades, researchers have experienced the difficulty of accurately carrying out NLP tasks. In the 1950s, researchers were optimistic about an imminent solution to the task of machine translation, while at the same time programming a human-level chess computer was considered an intractable problem (Weaver, 1955). Currently, chess computers are beating grand masters with ease, whereas commonly used machine translation systems such as Google Translate still have plenty of quirks. NLP tasks can be carried out both at the syntactic level, including tasks such as part-of-speech tagging and sentence parsing, and at the semantic level, including tasks such as named entity recognition, sentiment analysis, and word sense disambiguation. For a long time such NLP problems were mainly approached using hand-written rules and grammars. The ELIZA computer program, for example, managed to simulate a rudimentary conversation with a psychotherapist based on some basic pattern matching logic (Weizenbaum, 1966). However, with the advent of machine learning, statistical NLP has supplanted much of these laborious approaches. For example, using the Expectation Maximization algorithm, which was used to increase text classification accuracy by combining labeled and unlabeled data (Nigam et al., 2000), and using Conditional Random Fields, probabilistic models that can be applied to sequences of words or sentences (Lafferty et al., 2001). Even more recently, representation learning and deep

learning approaches have further improved the state-of-the-art in many NLP tasks (Bengio et al., 2013).

Within the medical domain, the subfield of clinical NLP focuses on processing clinical text, that is typically written by medical professionals (Chapman et al., 2011). Both the text and the tasks in clinical NLP differ from NLP in general. Clinical text, often written under time pressure, is characterized by ungrammatical sentence structures, by use of verbatim text, idiosyncrasies, and abbreviations, and by a domain specific vocabulary (Savova et al., 2010). Tasks include for example de-identification of text, matching text to ontologies, extracting clinical information, and identifying cohorts based on text. All clinical text we analyze in this research is written in Dutch. Although some impressive results have been achieved in English clinical NLP, these existing methods, tools, and platforms only have a limited applicability to our case. Throughout this dissertation, we therefore apply existing methods where possible, and develop our own methods for Dutch language where needed. In Chapter 4 for example, we design and implement a rule-based NLP method to automatically de-identify Dutch clinical text, while in Chapters 5 and 6 we use representation learning to improve text classifications, and in Chapter 7 we rely on clinical notes as a way to evaluate the nature of psychiatric patient subgroups.

1.4 Research Questions

There is great potential in analyzing EHRs to improve care, and its benefits and challenges have often been discussed. However, this type of research is still scarce in practice, both in general medicine and in psychiatry specifically. There is a lack of scientific methods that address the several challenges surrounding analysis of EHR data, and for this reason cases where EHR data are used to improve care are still very limited. In this dissertation, we therefore aim to contribute to improved knowledge discovery in psychiatric EHRs from a computing perspective. This constitutes identifying and addressing key technical, organizational, and ethical challenges of using EHRs for analysis, as well as applying knowledge discovery techniques to EHR data in order to obtain new insights that can improve care. We therefore pose the following main research question:

MRQ — How can data from Electronic Health Records provide relevant insights for psychiatric care?

Integrating data among different health care providers can already be considered difficult from a technical perspective, and this type of approach is further impeded by concerns regarding patient privacy and legislation. Our research is therefore scoped to analyzing EHR data *within* a health care organization. To answer the main research question, we pose six research questions, positioned in Figure 1.1. The first three research questions mainly pertain to prerequisites of using EHR data for analysis within a health care organization. Research questions four through six then build upon this foundation by identifying specific research problems in psychiatry, and investigating how analyzing EHR data can contribute to answering them.

RQ1 — How can health care professionals and data experts collaboratively perform exploratory analysis?

Data analysis is typically performed within an organization, such as a health care provider in the case of analyzing EHRs. Using knowledge specific to the organization and its data, also known as *domain knowledge*, is vital for performing successful and useful analyses (Tijsen et al., 2011). The data analyst, who is well versed in the technical part of analysis, initially lacks this knowledge. Practitioners on the other hand, including physicians, nurses, and other relevant staff, usually lack technical skills to perform analysis, but are an excellent source of domain knowledge. These two types of experts ideally need to collaborate, starting at the initial exploratory phases of a data analysis project. Failure to do so often leads to low-quality data analysis, and low likeliness of implementation of results (Brennan and Bakken, 2015). This research question investigates how data analysts and domain experts can collaboratively contribute in the various steps of a data analysis project, in order to obtain exploratory results that are supported by both.

RQ2 — What technical infrastructure can support reusing Electronic Health Record data for analysis?

In order to structurally use EHR data for analysis, several technical, organizational, and ethical challenges need to be addressed (Meystre et al., 2017). Technical challenges include data preprocessing and secure storage of data, organizational challenges include collaboration among researchers and repeatability of research, and ethical challenges include compliance with legal regulations and patient privacy (Safran, 2014). A technical infrastructure, consisting of appropriate hardware and software components that interoperate, can offer structural support when addressing these challenges (Danciu et al., 2014). Ex-

isting research has primarily focused on infrastructure to combine EHR data from multiple sites, on data management practices for clinical trials, or on one specific challenge that infrastructure can address. However, how a broad, unifying technical infrastructure that supports reusing EHR data for analysis within a health care organization should be designed is unknown. This research question investigates what the requirements for such an infrastructure are, and how an infrastructure that can satisfy all these requirements can be designed.

RQ3 — To what extent can clinical text written in Dutch automatically be de-identified?

In addition to structured data, EHRs typically contain information in free text format, for example in nurse and doctor notes, discharge letters, or treatment plans. Utilizing this data for analysis in addition to structured data often has substantial additional value (Harpaz et al., 2014). However, information that can either directly or indirectly identify a patient should be removed from these texts before they can be analyzed, in order to comply with legal and privacy requirements. Manual de-identification, by a human annotator, is prone to error, and moreover time-consuming and therefore expensive (Deleger et al., 2013). For this reason, automatic de-identification of clinical text before analysis takes place is a common approach that is already feasible in some languages (Meystre et al., 2014). This research question aims to investigate to what extent clinical text, i.e. text from EHRs, written in Dutch, can automatically be de-identified.

RQ4 — What Machine Learning techniques are useful for predicting inpatient violence?

In clinical practice, a lot of information in EHRs is captured in free text, which is less straightforward to analyze than structured data. Yet, these data can provide important information for various types of analysis, including predictive analysis (Chapman et al., 2011). By using a combination of Natural Language Processing (NLP) and Machine Learning (ML) techniques, in the past good results have been obtained in text mining problems. Recently however, novel Deep Learning techniques have been introduced, which have improved the state of the art in many NLP tasks (Goldberg, 2016). Most notably, the word2vec algorithm (Mikolov et al., 2013), and Recurrent and Convolutional Neural Networks, have contributed to this success. We hypothesize that inpatient violence incidents, including both physical and verbal violence from

patients, can be predicted by using free text from patients' EHRs as input. This research question investigates whether Deep Learning techniques have an additional benefit over classical machine learning techniques, when applied to predicting inpatient violence incidents.

RQ5 — How can automatic inpatient violence risk assessment using textual data contribute to the psychiatric practice?

After RQ4 shows which machine learning techniques are the most promising to predict inpatient violence incidents, we focus on violence risk management in psychiatric practice. Currently, violence risk assessment is typically performed using unstructured clinical judgment or using checklists (Singh et al., 2014). Both approaches have their drawbacks. Unstructured clinical judgment is subjective, and therefore prone to error. Checklists are known to have greater predictive validity on average but suffer from lack of robustness over different patient populations (Yang et al., 2010). RQ4 already gives a first impression whether violence risk assessment using clinical notes is feasible. However, an objective estimate of predictive validity and generalizability of trained models still needs to be established, along with the feasibility of an automatic assessment approach for the daily practice. This research question therefore aims to investigate how automatic inpatient violence risk assessment can contribute to better violence management within psychiatric care.

RQ6 — How can robust subgroups of psychiatric patients be identified based on Electronic Health Record data?

The identification of patient subgroups is an important process in many medical domains. In psychiatry, patient subgroups are mainly derived from the Diagnostic and Statistical Manual of Mental Disorders (DSM). However, critics of the DSM have pointed out that its diagnostic categories are heterogeneous, and that symptoms often cut across disorders (Cuthbert and Insel, 2013). Subsequently, data-driven approaches to identify patient subgroups using clustering algorithms have been proposed, but they have suffered from subjectivity in choosing a number of clusters and a clustering algorithm (Marquand et al., 2016). Cluster ensembles, i.e. combinations of multiple partitions, may be one method to obtain more objective subgroups. Therefore, this final research question investigates how applying cluster ensembles to EHR data can identify robust subgroups of psychiatric patients with clinically relevant characteristics.

Table 1.1: Overview of research methods in relation to research questions.

Method	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6
Expert interviews	•	•				
Focus group		•			•	
Prototyping		•	•		•	
Computational experiment			•	•	•	•

1.5 Research Methods

To provide answers to RQ1–RQ6, we will make use of several different research methods, both of a quantitative and a qualitative nature. In this section, the most important methods are briefly discussed in the context of this research, along with a description of how, and for answering which RQs, these methods were used. Although this list does not cover every research method that was employed in this dissertation, the most important ones are included. Additionally, Table 1.1 shows an overview of the research methods listed in this section, and in answering which RQs they were used.

1.5.1 Expert Interviews

Interviews are one of the most common types of qualitative research, where a participant's knowledge, belief, or opinion of a certain phenomenon is elicited, by means of face to face questions asked by a researcher (Britten, 1995). There are three main types of interviews that generate qualitative data: unstructured, semi-structured, and in depth interviews (DiCicco-Bloom and Crabtree, 2006). In the first case, questions emerge during the interview, and little or no questions are determined beforehand. In the second case the interview initially consists of predetermined, open-ended questions, and additional questions may emerge during the interview. In the third case, one or more topics are discussed in detail, based on clearly defined questions. The expert interview is a special case of the qualitative interview, where participants are very knowledgeable on the subject at hand. This type of interview is not without risks. For example, the knowledge of experts should not be considered absolute. Talking to experts however can be seen as an efficient way to obtain specialist knowledge, that is otherwise not easily obtainable (Bogner et al., 2009). In this research we use expert interviews to investigate RQ1 and RQ2, in both cases with the goal of exploring a relatively new topic. For this reason, we use interviews of the semi-structured type, that are then transcribed, and

analyzed using qualitative data processing techniques.

1.5.2 Focus Group

When knowledge, opinions, or beliefs of multiple participants need elicitation, instead of conducting individual interviews the researcher may choose to organize a focus group—essentially a group interview (Morgan, 1997). A moderator, who is not considered a participant in the focus group, guides the meeting by asking open-ended questions about a research topic of interest, to which the group's participants reply. One disadvantage of using focus groups is the increased distance between participant and interviewer, making it more difficult to ensure that all participants are heard equally (Morgan, 1998). Additionally, this can lead to loss of control on which aspects of the topic of research are discussed by the group. On the other hand, the main advantage of a focus group setting over individual interviews is that it allows group interaction, which can provide more direct insights into participants' similarities and differences in opinions on a topic (Kitzinger, 1995). In this research, a focus group is conducted during RQ2 and RQ5, both times to find participants' opinions about the research artefacts that are created.

1.5.3 Prototyping

In information systems, building a prototype of a model or system is an appropriate way to improve requirements, commitment from end-users, and quality of code (Beynon-Davies et al., 1999). Prototyping can be done in the early stages, eliciting or validating requirements, in the middle stages, confirming the behavior of a system, or in the late stages, investigating the operational system. In addition to a development tool, prototyping can also be used as a research method, to increase knowledge and understanding of a system (O'Leary, 1988). Building such a prototype can additionally be used to quantify the performance of a system. In this research, we use prototyping during RQ2, RQ3, and RQ5. In all cases we build a horizontal, working prototype, meaning that all functionality of the system is included in the prototype to study its workings, but not always to the level of the operational version of the intended system.

1.5.4 Computational Experiment

In the classical sense, an experiment is a procedure that is used to study the effect of modifying an independent variable on the dependent variable. In com-

puting however, the term experiment typically has a somewhat broader meaning than the controlled experiment (Tedre and Moisseinen, 2014). Zelkowitz and Wallace (1998) for example categorize experimental research in software engineering into observational, historical, and controlled methods. In the first case, data about new technology that is being studied is collected, while in the second case data about complete projects is collected, and in the third case multiple instances of an observation are studied. The first case is especially relevant in knowledge discovery, for example when an algorithm is applied to a dataset, resulting in empirical data that can be further analyzed or interpreted in a scientific context. In this research, we use computational experiments of the observational type in RQ3 and RQ6, in order to make statements about the computational model under study. In RQ4 and RQ5 we use experiments of the controlled type, studying the effects of changes in a computational model.

1.6 Dissertation Outline

The research questions RQ1–RQ6 presented in Section 1.4 are investigated in Chapters 2–7 of this dissertation, with each research question corresponding to a single chapter. Each of these six chapters are written as papers, published in proceedings of scientific conferences or in scientific journals. In many chapters EHR data is used, an overview of the type of data and number of records that were used in each chapter are shown in Table 1.2.

Chapter 1 — Introduction describes the motivation and objectives of this research, along with the medical research domain of psychiatry, and the computing research domains relevant to this dissertation. We introduce the main overarching research question, six specific research questions, and research methods that are used to investigate these questions. Finally, the dissertation outline, in this section, describes how the remainder of this dissertation is structured.

Chapter 2 — Expert Sessions for Knowledge and Hypothesis Finding explores how data experts and health care professionals can work together in finding new knowledge and hypotheses for further research. For this purpose we propose the CRISP-IDM, where the I stands for Interactive, a process for collaboration between the two using data visualization. We organize 19 sessions where data experts explore EHR data together with various health care professionals, based on a selection of elicited research themes and initial questions. We show that organizing these data-driven sessions yields interesting

Table 1.2: Overview of datasets used in each of the individual chapters.

Chapter	**Research Question**	**Dataset**
2	RQ1	5,800 DSM diagnoses 6,500 treatment plans 22,000 medication prescriptions 13,000 checklist responses 5,400 admissions 150,000 nurse notes 1,200 violence incident reports
3	RQ2	n/a
4	RQ3	1,200 nurse notes 1,200 treatment plans
5	RQ4	2,251 admissions 25,942 clinical notes 1,248 violence incident reports
6	RQ5	3,189 and 3,253 admissions 37,853 and 81,641 clinical notes 962 and 652 violence incident reports
7	RQ6	1,098 Youth Self Report questionnaires 665 DSM diagnoses

new knowledge and hypotheses for further research, much of which were initially not imagined. This chapter has been published in the Computational and Mathematical Methods in Medicine journal (Menger et al., 2016).

Chapter 3 — Infrastructure for Supporting Reuse of EHR Data investigates how infrastructure, consisting of various software and hardware components, can support analysis of secondary EHR data within a health care organization. For designing such infrastructure, we propose the Capable Reuse of EHR Data (CARED) framework. We first conduct exploratory interviews with experts of relevant disciplines in the UMCU, and then identify nine requirements that modern infrastructure should address. We propose the CARED framework as a specification of the Data Warehouse model of Inmon (2002), adapted to fit these requirements. After evaluating the framework with respect to the requirements, we describe how it was used to guide design and implementation of infrastructure in the Department of Psychiatry of the UMCU. This chapter has been published in the proceedings of the International Conference of Health Informatics (HEALTHINF) (Menger et al., 2019a).

Chapter 4 — De-identification of Dutch Medical Text introduces the De-identification Method for Dutch Medical Text (DEDUCE): a pattern matching method for automatic de-identification of clinical text written in Dutch. We identify a number of Protected Health Information (PHI) categories that are present in clinical text in patients' EHRs, such as names of patients and family members, locations, names of health care organizations, and dates. We then develop and evaluate a pattern matching based method for annotating and removing information in these categories. We achieve a recall of 0.964 for person names, and a micro-averaged F_1-score of 0.862 over all categories, in line with state-of-the-art results in other languages. This chapter has been published in the Telematics and Informatics journal special issue on Applied Data Science in Patient-centric Health Care (Menger et al., 2018b).

Chapter 5 — Predicting Inpatient Violence Incidents applies several machine learning techniques to a classification task, with the objective to predict violence incidents during admission. The dataset, obtained from patients' EHRs, consists of 2,521 psychiatric admissions, reports about violence incidents, and clinical notes written by practitioners up to the first 24 hours of admission. We apply several methods for representing the clinical notes, both based on bag-of-words and based on representations learned using a large corpus of clinical text. Subsequently, we employ several classification models, including methods such as Naive Bayes and Support Vector Machines, and neural models

such as Recurrent Neural Networks. Results show that state-of-the-art techniques obtain a relatively small, but consistent improvement in performance over the more traditional methods. This chapter has been published in the Applied Sciences journal special issue on Data Analytics in Smart Health Care (Menger et al., 2018a).

Chapter 6 — Violence Risk Assessment using Clinical Notes investigates to what extent automatic violence risk assessment can benefit the psychiatric practice. We obtain datasets from the EHRs of two independent treatment sites, again consisting of psychiatric admissions, reports of violence incidents, and clinical notes. Partially based on the results of the previous chapter, we perform a rigorous evaluation of trained machine learning models applied to these datasets separately, and exchange trained models to assess their generalizability. Our findings show that machine learning models can assess violence risk with good predictive validity, in the same range of existing assessments. This finding highlights the potential value of analyzing routinely written clinical notes. However, predictive validity of pre-trained models is significantly lower. This chapter has been published in the JAMA Network Open journal (Menger et al., 2019c).

Chapter 7 — Identifying Psychiatric Patient Subgroups investigates how cluster ensembles, i.e. combinations of multiple clustering algorithms, can help increase robustness and reduce variation in reported stratifications of psychiatric patients. We first propose a Meta Algorithmic Model (MAM) to guide researchers in applying cluster ensembles to their specific (medical) domain. We then apply our MAM to a dataset of 1,098 Youth Self Report questionnaires, describe and evaluate the clusters we obtained, and assess their relations to several clinically relevant variables including DSM diagnosis. This chapter has been published in the proceedings of the Conference on Artificial Intelligence in Medicine (AIME) (Menger et al., 2019b).

Chapter 8 — Conclusion provides answers to the research questions, based on the investigations in the individual chapters. We furthermore describe the limitations of this research, directions for further research, and close with personal reflections.

2 | Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding

The surge in the amount of available data in health care enables a novel, exploratory research approach that revolves around finding new knowledge and unexpected hypotheses from data instead of carrying out well-defined data analysis tasks. We propose a specification of the Cross Industry Standard Process for Data Mining (CRISP-DM), suitable for conducting expert sessions that focus on finding new knowledge and hypotheses in collaboration with local workforce. Our proposed specification that we name CRISP-IDM is evaluated in a case study at the Department of Psychiatry of the University Medical Center Utrecht in the Netherlands. Expert interviews were conducted to identify seven research themes in the department, which were researched in cooperation with local health care professionals using data visualization as a modeling tool. During 19 expert sessions, two results that were directly implemented, and 29 hypotheses for further research were found, of which 24 were not imagined during the initial expert interviews. Our work demonstrates the viability and benefits of involving work floor people in the analyses, and the possibility to effectively find new knowledge and hypotheses using our CRISP-IDM method.

This work was originally published as:

Menger, V., Spruit, M., Hagoort, K., and Scheepers, F. (2016). Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Computational and Mathematical Methods in Medicine*, 2016:1–11

2.1 Introduction

With the increase of the amount of data that is collected and stored in many different fields in this digital age, the amount of data that is being collected in health care has also increased enormously over the past years. EHR software has become more widespread in hospitals and other health care institutions, both worldwide and in Netherlands (DesRoches et al., 2013; Michel-Verkerke and Spil, 2002). This central approach to patient data management enables using patient data for clinical research purposes through various data mining techniques (Murdoch and Detsky, 2013; Raghupathi and Raghupathi, 2014).

Where Randomized Controlled Trials are the current gold standard in mental health research, the sample size and lack of selection bias of a data analysis approach can offer a novel and substantial contribution to mental health research by discovering complex interaction patterns of cause and outcome, resilience and vulnerability, and risk and outcome (Torous and Baker, 2016). In previous studies, data driven research has led to new scientific knowledge, improvement of patient treatment, and reduction in administrative load, or financial savings, both in general health care (Koh et al., 2005) and in mental health care (Kraemer and Freedman, 2014; Lahti, 2016; Passos et al., 2016).

Although these approaches are undeniably beneficial, they are hypothesis driven, and therefore do not necessarily exploit the full potential of data driven research. Data analysis enables the possibility of uncovering relations, patterns, and trends that were previously neither expected nor hypothesized (Khoury and Ioannidis, 2014; Kitchin, 2014; Oquendo et al., 2012). Generated hypotheses can subsequently lead to replication studies with higher probability of success, because of their groundedness in the data (van Helden, 2012). This exploratory approach comes with its own challenges, yet it also appears to have great benefits for the health care domain (Raghupathi and Raghupathi, 2014).

Another possible benefit of this exploratory approach is the ability to involve local professionals and patients in the analysis process. Where currently much of the data analysis literature in health care focuses on the technical aspects of the transition to a more data driven standard (e.g. Chawla and Davis, 2013) transforming the workforce mindset and business processes in daily practice is a challenge that is not to be underestimated (Wang et al., 2014). The staff that has to work with the outcomes of a data analysis project is usually unfamiliar with the concept of data analysis, which creates a gap between domain experts and the technical staff performing the analysis (Meulendijk et al., 2013; Tijsen et al., 2011). Moreover, not involving clinical practitioners in the

project will lead to them feeling surpassed, which will lead to a failure to adopt the technology, and thus a failure of the project. A high amount of interaction with the local workforce during a data analysis project will mitigate this problem, by constantly requiring their input in determining relevant topics and outcomes (Brennan and Bakken, 2015). Cooperation with domain experts will both strengthen the analysis, and pave the way for an easy implementation of results that are eventually discovered.

In an effort to benefit from the opportunities of data driven research, and to simultaneously tackle the challenges of overcoming the gap between domain experts and technical staff, we report on an interactive data analysis project that focuses on collaborative knowledge and hypothesis finding in the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in Netherlands. This work contributes to the field of data analysis in mental health care by exploring the benefits and limitations of a predominantly non-hypothesis driven approach that actively involves the local workforce in the process, and in that way helps mental health care institutions transition to a more data driven practice. This specific interactive data analysis approach that we have pursued has not been applied before in the context of mental health.

2.2 Materials and Methods

To investigate the possibilities and limitations of an interactive data analysis project as described in the introduction, we conducted a single exploratory case study (Yin, 2009) at the Department of Psychiatry of the UMCU. The goal of the case study is to explore the benefits and limitations of a data analytics project that:

1. focuses on finding new knowledge and hypotheses, and
2. involves local health care professionals in every step of the process.

After researching and comparing the Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000), Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996), Sample Explore Modify Model Assess (SEMMA) (Azevedo and Santos, 2008), and 3 Phases Method (3PM) (Vleugel et al., 2010), CRISP-DM was selected to structure the case study. CRISP-DM is a well-defined process that is widely adopted in the industry (Onwubolu, 2009). Moreover, it emphasizes the organizational part of data mining, which is in line with our project goals. CRISP-DM consists of the following six steps:

domain understanding, data understanding, data preparation, modeling, evaluation, and deployment.

For making the CRISP-DM suitable for doing the exploratory data analysis that also involves health care professionals in every step of the process, we propose the following three modifications. First, we aggregate the modeling and evaluation phases into one iterative phase, that requires collaboration with domain experts. To enable this, data visualization will be used as a modeling tool, allowing participation for those unfamiliar with data analysis. Second, we distinguish between the general and the specific preparation phases, in which, respectively, generic preparation tasks, and tasks that are revealed during the exploratory modeling and evaluation phase are carried out. Third, an optional inferential analysis step is added that in many cases will be necessary to be able to bring exploratory analysis results or generated hypotheses to the daily practice with sufficient confidence. An overview of the approach process is depicted in Figure 2.1, the left column showing the different phases and their relation, and the right column showing the most important goals of each phase. This overview can be seen as a specification of CRISP-DM to make it applicable to doing exploratory and interactive data analysis, which we named CRISP-IDM: Cross Industry Standard Process for Interactive Data Mining. While we have applied this approach in the context of mental health care, it can easily be applied to other domains as well, since it is generic, and relies on local domain experts and data sources. From April 2015 through November 2015, the case study was implemented along the phases of our proposed CRISP-IDM method.

2.2.1 Domain Understanding

The case study was conducted at the Department of Psychiatry of the UMCU. It consists of four units that specialize in severe affective disorders, psychotic disorders, developmental disorders, and urgent care. There are roughly 60 inpatient treatment positions. Combined with outpatient treatment, this results in approximately 2,000 unique patients annually. Relevant clinical staff consists mainly of psychiatrists, psychologists, nurses, and additional staff that directly support treatment such as therapists and social workers. The department has a secondary care function regionally, as well as a tertiary care function nationally. This entails the presence of a diverse population, including patients that have been referred by general practitioners or other primary care institutions, patients that require urgent care, and patients with more complex symptomatology. As a university medical center, the UMCU is assigned to conduct research, which makes this center suitable for carrying out

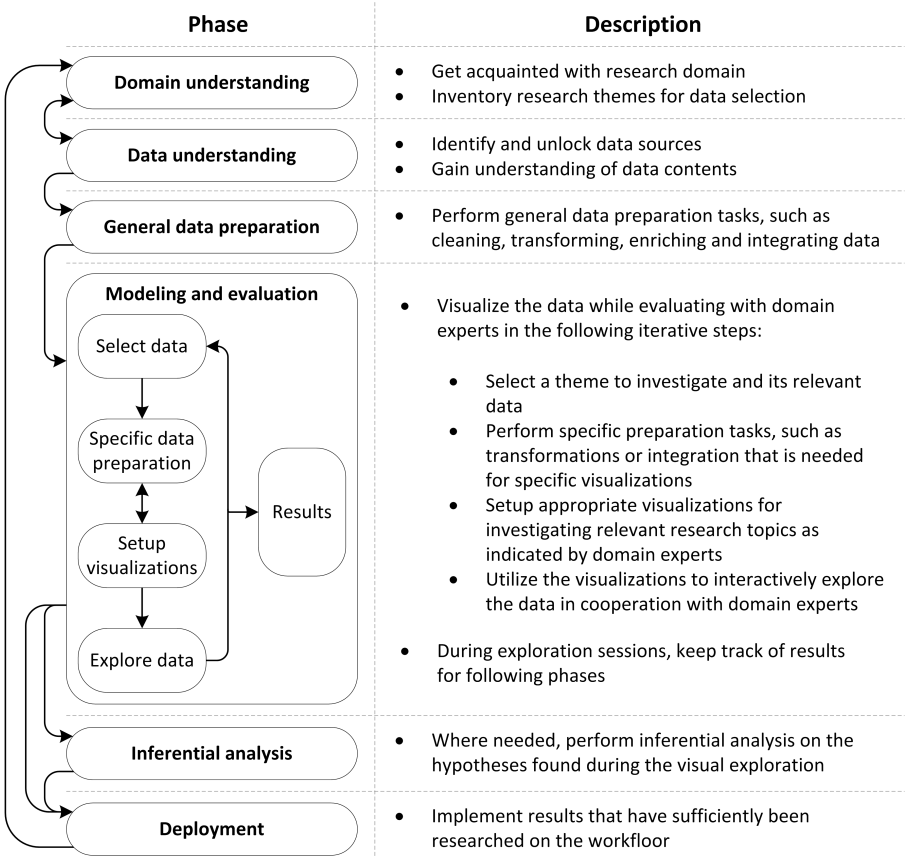


Figure 2.1: Overview of the CRISP-IDM method: a CRISP-DM based process for interactive data mining.

the case study.

In order to become acquainted with the department, semistructured interviews were conducted with six different representatives. This included a psychiatrist of each of the four psychiatry units, a board level psychiatrist, and a psychologist. According to Yin (2009) semistructured interviews are appropriate when exploring a new topic. The initial questions were open questions pertaining to the data that is recorded within the department, and possible research directions with this data. Interviews were transcribed afterwards. Although the data mining project is explicitly non-hypothesis driven, that is, not intended to answer specific research questions, interviews with health care professionals resulted in a clear mapping of specific themes that are needed for selecting relevant data sources and general data preparation. From these six interviews, a total of 28 topics were identified using Inductive Content Analysis (ICA) (Elo and Kyngäs, 2008). Subsequently, this list of topics was further categorized into seven themes, again using ICA. In a follow-up meeting with five of the six interviewed health care professionals, both the individual topics, and the higher level themes were rated on their relevance, resulting in a prioritization of both. The five most relevant topics of the 28 are displayed in Table 2.1, and the themes they are categorized into in Table 2.2. Once again, the reason for identifying the topics is not that they must be researched during the project, but for providing general insight into the type of analyses that can be conducted; for guiding the analysis only the higher level themes were used. The prioritization was determined by the number of times a topic was judged to be of clinical relevance by experts under constrained selection, and for the themes in Table 2.2 this was conducted in the same way.

The themes 2 and 3 in Table 2.2 show a direct relation to the daily practice of health care professionals. The admission and dismissal theme mostly pertains to questions regarding the length of admission and the likeliness of readmission, while the aggression theme concerns questions about aggression incidents that occur with local inpatients. The context factors (theme 1) and patient referrals (theme 4) themes however mostly relate to the part of the care process that occurs outside of the UMCU, for example, the social and economic background of a patient, and the continuing of a treatment in other institutions. Themes 2 and 3 therefore require data that is locally generated and stored, and themes 1 and 4 require external data that can be connected to the local data.

Table 2.1: The five identified topics with the highest priority, along with the theme they are categorized to. Note that these are not questions that must be answered during the project, but serve as a way to become acquainted with the department as a case study domain. Abbreviations: ROM = Routine Outcome Monitoring.

Topic	**Theme**	**Priority**
What are relations between the different ROM scores (or between specific sections of the ROM scores), and can they predict length of treatment?	ROM	1
Do medication prescription and change in medication influence the length of admission and the likeliness of readmission?	Medication	2
Can aggression incidents in inpatients be predicted?	Aggression	3
In what way are patients referred between for example general practitioners, secondary care institutions, and the UMCU?	Patient referrals	4
What are descriptive statistics of aggression incidents, i.e. in what location, at what time, and with which members of staff do they occur?	Aggression	5

Table 2.2: The seven themes that were derived from the topics that were mentioned in expert interviews (Table 2.1), and their priority.

Theme	**Priority**
Context factors	1
Admission & dismissal	2
Aggression	3
Patient referrals	4
Routine Outcome Monitoring	5
Medication	6
Other	7

2.2.2 Data Understanding

Researching the identified research themes in Table 2.2 requires selection of relevant data sources, gaining access to these data sources, and unlocking their relevant data. Within a health care institution multiple possible data sources usually exist, such as an EHR, lab measurements, imaging data, and other databases that contain relevant information (Weber et al., 2014). Externally, census data, geographical data, and data gathered by other care institutions contain information relevant for analysis. These external data sources are important for doing analysis concerning the environmental factors that interplay with development and treatment of symptoms, as well as comparing the patient group of the UMCU with healthy citizens or patients in other care institutions.

Lab measurements and imaging data were omitted due to few mentions during the domain understanding phase, and limited amounts of available data on our patient population. It was not feasible to obtain data from other care institutions within the duration of the project, due to privacy constraints and limited resources. The ultimately accessible data sources, and their data are listed in Table 2.3. The various entities are representative of all important care process aspects, and allow performing analyses in all themes except patient referrals. For the sake of enabling exploratory data analysis, all variables were initially included for each of the datasets. The purpose of the exploratory analysis is finding new and unexpected relations or patterns, so naturally no variables should be excluded at the outset.

From the EHR, several data entities were available:

1. All patients receive a diagnosis according to the standard Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV) classification system on four axes: (1) primary diagnostic for treatment, (2) personality disorders, (3) medical or physical disorders, and (4) psychosocial and environmental factors.

2. When starting treatment, a treatment plan is written by a health care professional, in which the type and duration of treatment are described, along with free text fields describing the symptoms and background of a patient.

3. During treatment, patients may have one or more medication prescriptions, which include the type, dose, and frequency of medication, the period in which the medication is prescribed, and possible mutations.

4. For the purpose of monitoring the state of a patient, Routine Outcome Monitoring (ROM) is performed by scoring a questionnaire or metric on a certain time interval. Available ROM methods include Health of the Nation Outcome Scales (HONOS), Kennedy Axis V, Child Behavior Checklist (CBCL), and the Global Assessment of Functioning (GAF). These all measure another aspect of the well-being of a patient, and are all taken at different moments in the care process with different time intervals.

5. Admission information comprises, for example, the date of admission and dismissal, to which unit the patient was admitted, and the type of admission.

6. Free text reports are written by psychiatrists and nurses about their admitted patients during each of the three shifts on a day. These unstructured text fields contain information on the daily well-being and activities of a patient.

When an aggression incident occurs, mandatory reporting takes place in a separate Incident Reporting System (IRS). For each incident, several structured variables are recorded, such as the patients and staff involved, the location and time of the incident, and type of aggression. Additionally, in free text variables information about the events leading up to the incident, and the incident itself, are captured.

Externally, open data from the Dutch national census bureau was acquired, such as statistics about the average income, urbanization, and type of homes in the living environment of the patient. From a geographical source, more detailed data about the amount of green space in the direct vicinity of patients was obtained.

2.2.3 Data Preparation

Since the available data is stored in a way that is not intended for doing research, preprocessing on the data was necessary to convert it to an appropriate format for exploratory analysis. The most important tasks in preparing the data are transforming, cleaning, integrating, reducing, and discretizing the data (Zhang et al., 2003). Since in the exploratory, non-hypothesis driven type of approach not all data preparation steps are known in advance, at first only the general tasks that can be carried out without knowing specific modeling goals were performed. We therefore call this phase the general data prepa-

Table 2.3: All acquired data entities, and their sources, type, structuredness, and number of records (rounded to the nearest hundred or thousand).

Source	**Data entity**	**Type**	**Structured / unstructured**	**No. records**
EHR	1. Diagnosis	Categoric	Structured	5,800
	2. Treatment plan	Categoric, textual	Both	6,500
	3. Medication prescriptions	Categoric, numeric	Structured	22,000
	4. Routine Outcome Monitoring	Numeric, textual	Both	13,000
	5. Admission information	Categoric	Structured	5,400
	6. Daily reports	Textual	Unstructured	150,000
Incident Report System	7. Aggression incident reports	Categoric, textual	Both	1,200
External	8. Census data	Numeric	Structured	21,000
	9. Geographic data	Numeric	Structured	5,000

ration phase, and iteratively carry out the specific data preparation in the modeling and evaluation phase.

First of all, the data transformation consists of simple tasks such as parsing date variables, and changing variable names. Many categoric fields do not contain data that is understandable for a user by default but instead contains codes or abbreviations; these fields were transformed to be user readable. In this stage of the process, no transformations were done with regard to incomplete data, because the modeling software that is discussed in Section 2.2.4 is designed to handle this. For textual data, several transformations such as stemming and stop word removal were applied, to achieve both data reduction and noise reduction.

After transformation, data was cleaned, for example, by removing redundant variables and duplicate records. This also includes identification variables and metadata that are meaningful in the system the data was sourced from, but not in the context of analysis.

Finally, to be able to make statements about concepts that are captured in different datasets, datasets that were often mentioned in relation to each other in the expert interviews were integrated. Additionally, in this step the data was enriched by using open data from both the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the Anatomical Therapeutic Chemical Classification System (ATC). This allows easier patient selection by looking at diagnoses and medication as a hierarchy, for example, distinguishing the anatomical main group level and the chemical substance level of a drug.

2.2.4 Modeling and Evaluation

To be able to involve the workforce in the modeling and evaluation phases, as well as doing research in an exploratory way, a different approach than the traditional one is required. In a usual setting, a team of data analysts or technical experts receive a specific problem that is posed by domain experts, use technical or statistical modeling software to answer the problem, and then collaborate with the domain experts again to evaluate, and possibly refine the models. This setting is not appropriate for our proposed approach, both because our approach is not aimed at answering specific questions but at exploring the data to find new knowledge and hypotheses, and because our approach aims to include health care professionals in every step of the process. Therefore an interactive data visualization tool is used to model the data. This tool enables direct feedback from health care professionals who can be present and directly participate in the modeling process in an approachable way.

The visualization tool supports several types of visualizations, ranging from

basic to advanced. Basic visualizations, for example, include scatter plots, histograms, and map plots. More advanced examples are network visualizations, bar charts, trend lines, and word walls that enable visualization of text variables. Furthermore, the tool supports a wide range of selection and drilling down options that easily enables zooming in on specific parts of the patient population. Most notably, the tool has an interactive modus operandi, which supports real time updating of selection sets in multiple visualizations, and therefore suits well with our goal of domain expert participation with direct feedback. This method excels in its visual interface and interactive nature, as opposed to traditional data mining tools which usually center around programming code and textual or numeric output.

The modeling and evaluation phase is done in weekly sessions that require the presence and collaboration of both health care professionals and technical staff. The health care professionals are needed to guide the analysis and find new interesting visualizations with their expertise of the domain, and the technical staff is only needed to facilitate the analysis by introducing the data and visualizations used, and operating the software. Note that the exploration was guided by the domain experts, and only facilitated by the technical staff, since they lack knowledge of the psychiatry domain. The weekly sessions were held 3–4 consecutive times for each of the identified themes in Table 2.2. The process consists of five iterative steps (Figure 2.2):

1. In the data selection step, initially a small set of data relevant to the current theme was selected. In next iterations, additional data was added to this set, whenever indicated to be relevant by the health care professionals.

2. In the specific preparation step, preparation that was not done in the general step is carried out where necessary for specific purposes. This includes integrating different data entities that were not linked in the source data, and data transformations or derivation of variables that were needed for specific visualizations requested by present health care professionals.

3. The visualization setup step constituted loading the prepared data into the visualization tool, and initially creating simple, descriptive visualizations of the data at hand, such as the number of admissions over time, distributions of the diagnoses, and most prescribed medications. In next iterations, more complex visualizations based on practitioner input were added.

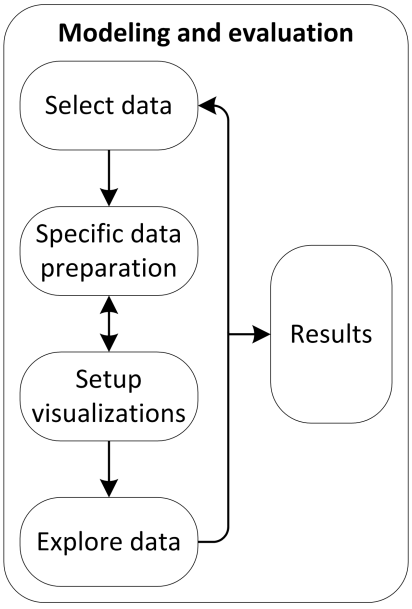


Figure 2.2: The modeling and evaluation phase process in CRISP-IDM.

4. During the exploration step health care professionals explored the data, supported by technical staff. The exploration was guided by the health care professionals, for example, by selecting patient population subgroups based on their different characteristics, comparing trends over time, and looking for patterns in the visualizations. The technical staff facilitates this analysis, by performing the actions that are requested by the health care professionals, and by helping interpret the visualizations. The collaboration during the session enables a creative process that leads to new ideas for visualizations, and the relating of new concepts and datasets by domain experts. These were explored in the next exploration cycle, because they often required additional data and preparation that was not yet performed in steps 1–3. Steps 1–4 are therefore iterative, since not all visualizations are initially known; they are a result of the creative process of looking at the data in collaboration.

5. The constant visual feedback of the modeling process enabled direct observation of which visualizations depict interesting results, such as possible correlations, textual terms that stand out or trends in bar charts. The products of steps 1–4 are these kinds of results that require action, that is, relations that can visually be confirmed and require further research. These observations are the direct result of the modeling and evaluation phase, and were noted for following phases of the CRISP-IDM process.

As an example, during the iterations of the aggression theme, a peak in aggression incidents was found to exist on the fifth day of admission. Initially, a small set of data was selected for the aggression theme, including the aggression and the diagnosis datasets. During the first iteration, only some descriptive statistics on aggression were visualized, such as the number of incidents per unit, and the most occurring types of aggression (e.g. verbal, physical). Experts then indicated that they were interested to see if there was any change in the number of incidents over time. This visualization was added in the next iteration but did not show any interesting results. Another expert wondered if it was more likely for an incident to happen at the beginning of the end of an admission. In the next iteration, admission data was therefore added, and integrated with the aggression data. Visualization showed a peak in incidents on the fifth day of admission (Figure 2.7), which domain experts did not expect, and judged to be unknown. This peak in incidents was therefore noted as a result for further inspection during the following phases. Other results, and their strengths, are further elaborated upon in Section 2.3.

In a period of three months, a total of 19 interactive exploration sessions with two or three domain experts, and at least one technical expert present

Table 2.4: For each function, the total number of unique people involved, and the total number of attendances in the interactive exploration sessions.

Main function	No. persons	Attendances
Psychiatrist	4	11
Psychologist	1	4
Nursing staff	4	9
Board	2	7
Policy maker	4	6
Other	3	4
Total	18	42

were conducted. A total of 18 health care professionals were involved in exploring the data, and each expert attended 2.3 times on average. A breakdown of the number of people and attendances is visible in Table 2.4, where it can also be verified that professionals from each part of the department process were included in the sessions.

2.2.5 Deployment

The deployment phase typically focuses on implementing the results that were obtained in the modeling phase on the work floor. This part of the process was in our case not very distinct from a typical data mining project, since results that were obtained and confirmed needed to be transformed to daily work practice. This concerns both results that were judged to be strong enough during the visual modeling, as well as results that were additionally tested using inferential analysis. The process that describes how results are deployed is depicted in Figure 2.3.

Initially, a selection of results that can contribute to daily work practice was determined with the help of medical management, including the board and the leading psychiatrists of each of the four units in the department. After a first rough selection is made, these results were discussed with relevant nursing staff and local researchers or policymakers. Incorporating both board and management level professionals as well as daily workforce ensured that eventually implemented results are widely supported in the department. Discussing the results with workforce can lead to the agreement that an outcome is not suitable for implementation, in which case no further action is required, or a positive response. In this case, the result is either ready for implementation, or needs further research, such as connecting the result to relevant literature,

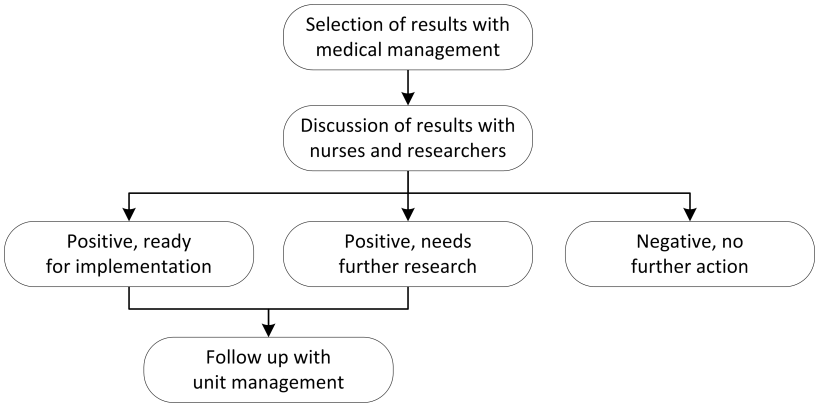


Figure 2.3: The deployment phase process in CRISP-IDM.

or exploration of other data sources or the EHR to investigate the result on a higher level. Depending on the type of result, further actions were determined, and assigned in a follow-up with the management of the relevant unit(s).

2.3 Results and Discussion

Two of the results that were found during the modeling phase were implemented on the psychiatry work floor:

1. The Kennedy Axis V, a specific Routine Outcome Monitoring method that is used in all four units of the department, scores the well-being of a patient in eight areas on a 0–100 scale, and additionally enables nurses to enter a note or explanation in a text field. The numeric scores are supposed to reflect the current state of a patient's well-being on regular intervals (e.g. weekly) during an admission. Data analysis however shows that for seven of the eight subscales the score remains almost constant during an admission (Figure 2.4), thus invalidating the need to score it on a regular basis. As a result, the numeric scores are filled in less frequently, yielding time savings, and lower administrative load at the work floor.

2. The nurse reports that are written about admitted patients during each of the three daily shifts were shown to be of varying extensiveness, both

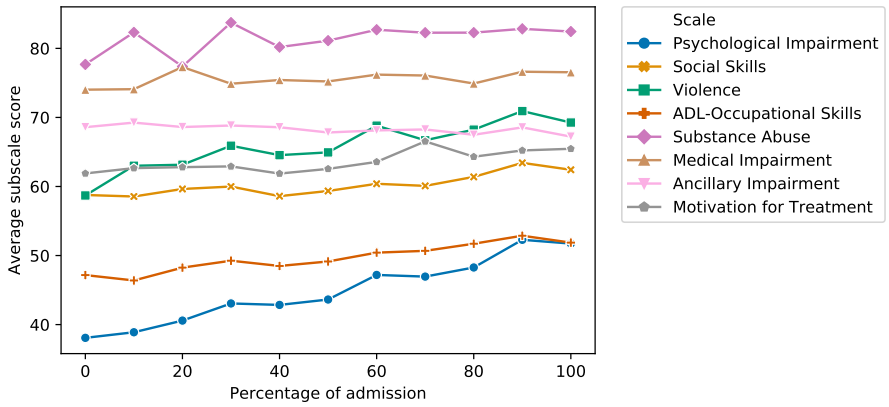
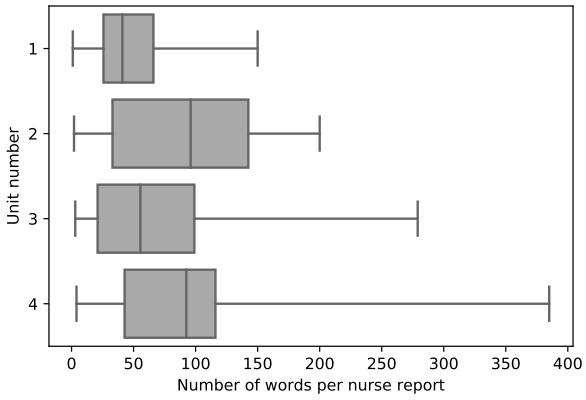


Figure 2.4: The average score of the eight Kennedy Axis V subscales over time, with the x-axis representing the percentage of the total admission length (429 admissions, 2,531 Kennedy Axis V reports). It can be seen that except for the Psychological Impairment subscale, the variation during the admission is minimal. Domain experts indicated that the amount of variation that can be seen does not justify scoring the Kennedy Axis V on a regular basis. Abbreviations: ADL = Activities of Daily Living.

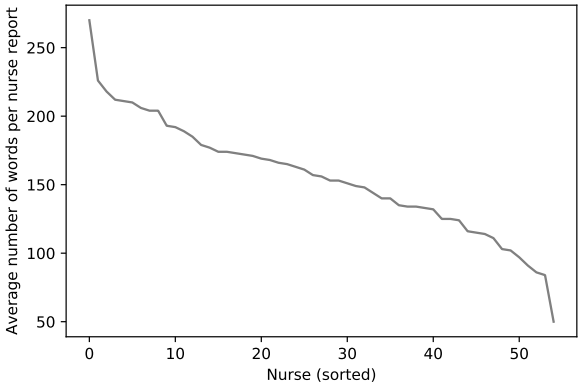
among different nurses, and among different units within the department (Figure 2.5). Among health care professionals there was a broad agreement that good reports are concise, so that the writing is not time consuming, and so that readers of the reports can quickly recognize key points. This has resulted in a new guideline for writing nurse reports, again leading to a decreased administrative burden for nurses.

In both cases, the fact that the people who have to work with the outcomes of the project aided in finding these outcomes paved the way for an easy implementation of these results.

During the modeling and evaluation phase some limitations of the interactive visualization software came to light as well. First, although the interactive visualization software used allows easy participation for nontechnical staff, this feature comes with a trade-off in methodological rigor. For example, the software does not allow statistical testing, or more sophisticated data analysis tasks. This kind of advanced analysis is more challenging to conduct in the presence of health care professionals because it usually centers around numer-



(a) Report length over units



(b) Report length over nurses

Figure 2.5: On top (a), a box plot for each of the four units can be seen that shows the variation of the number of words in a nurse report (140,719 reports). It can be seen that there are substantial differences among the units, for instance by looking at the median number of words in a nurse report: for Unit 1 this is around 50, but for Unit 2 this is around 120. On the bottom (b), the average status report length per nurse for Unit 3 is depicted (33,418 reports). Between different nurses, large differences can occur. Other units show very similar distributions. In both images, only nurses that wrote at least 50 statuses were included in the analysis.

ical output instead of visualization, which is more difficult to understand for clinicians. At times a correlation, pattern, or trend appeared to exist in a visualization, yet a decisive answer cannot be given without reverting to more traditional statistical software.

Second, while reverting to the statistical software can in many cases determine the strength of a result with a correlation coefficient or a p value, the fact that many other relationships between variables have been discarded without any testing makes this an unusual instance of the multiple comparisons problem. Even if a statistically significant relation is found, the fact that many relations have been manually examined without any testing weakens the result. Although this effect is usually stronger in machinal computation of many correlations, it is definitely present here, and it prohibits simply viewing the outcome of statistical test as a direct result. The exploratory character of the expert sessions therefore limits the ability to conclusively verify results, and in many cases additional research, for example, on other similar datasets, is required.

For the reasons mentioned above, other results that have been found in the data are classified as hypotheses for further investigation, rather than direct results. Put differently, the exploratory outcomes provide the input for inferential analysis. The two implemented results differ from these hypotheses in that the data describes the entire population, and a statistical method is not indispensable to validate this kind of result.

2.3.1 Hypotheses for Further Research

From the result step in the exploration process, a list of 29 hypotheses and topics for further investigation were identified. As opposed to the list of topics identified in Table 2.1, all of these assertions have some basis in the data. Of the 29 hypotheses, 5 are marked as directly related to one of the topics in Table 2.1; the other 24 have explicitly been found while exploring the data. This makes the proposal of hypothesis finding using interactive data analysis a success, since a large part of the found hypotheses are distinct from the questions that were known before looking at the data. Figure 2.6 shows a breakdown of the number of found hypotheses per theme. Note that the Patient Referral theme was abandoned in the data understanding phase since obtaining data from other institutions was not viable within the duration of the project. It can be seen that most hypotheses pertained to the aggression theme, followed by context factors, admission and dismissal, Routine Outcome Monitoring, and medication themes with a roughly equal number of hypotheses each.

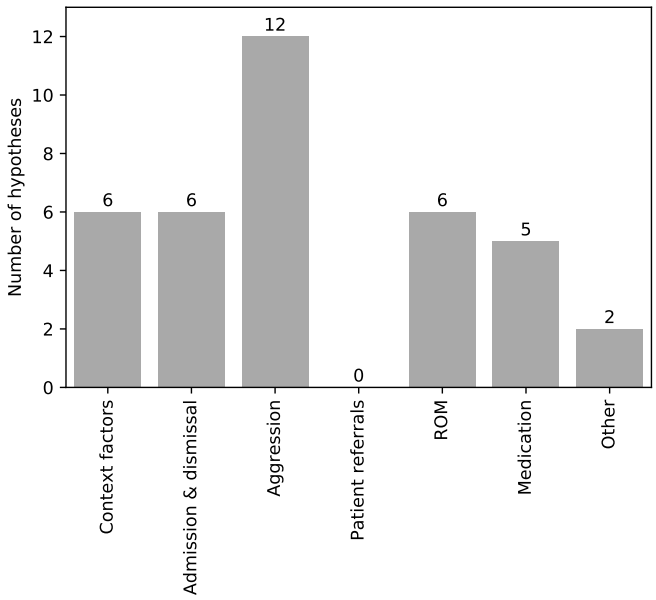


Figure 2.6: Number of hypotheses for further research found per theme, note that multiple themes can apply to one hypothesis. Abbreviations: ROM = Routine Outcome Monitoring.

Table 2.5: Five of the hypotheses for further research selected by domain experts, in random order. Note that these hypotheses have some basis in data, but need further research to turn into results. The second and fifth hypothesis are visualized in Figures 2.7 and 2.8. Abbreviations: ADHD = Attention Deficit Hyperactive Disorder.

Hypothesis	**Theme**
There exists a positive relation between season of admission and length of admission (longer admissions during winter)	Admission
A peak in aggression incidents occurs on the fifth day of admission (Figure 2.7)	Aggression
There exists a relation between aggression incidents and wearing of of medication effects in patients diagnosed with ADHD	Aggression, medication
There is an absence of a relation between amount of green space in patient environment and likelihood of developing a disorder	Context factors
There is a negative relation between economic status of living environment and length of admission (Figure 2.8)	Admission, context factors

A top five of hypothesis for further investigation selected by health care professionals is displayed in Table 2.5. It is important to note that these hypotheses have a basis in data but for reasons mentioned above are not to be interpreted as direct results; for this additional research is required.

2.3.2 Process Evaluation

Carrying out the case study along our CRISP-IDM method generally went well, although some practical difficulties were experienced as well. Some of the most important findings for each of the phases are listed below:

1. Domain Understanding. Conducting expert interviews and identifying the seven research themes generally succeeded in getting acquainted with the research domain. The importance of this step cannot be underestimated, both because it forms the basis for the data selection, and thereby directly influences the successfulness of the project, and because it is the first step in interacting with local workforce.

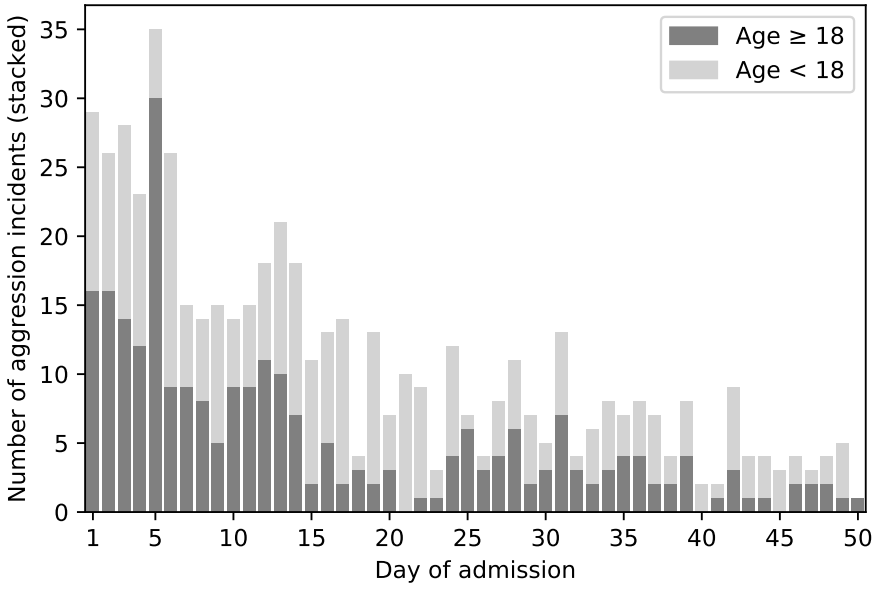


Figure 2.7: The total number of aggression incidents for the first 50 days of all admissions is visible (631 incidents). It can be seen that a peak occurs at day five, especially in adult patients (dark blue).

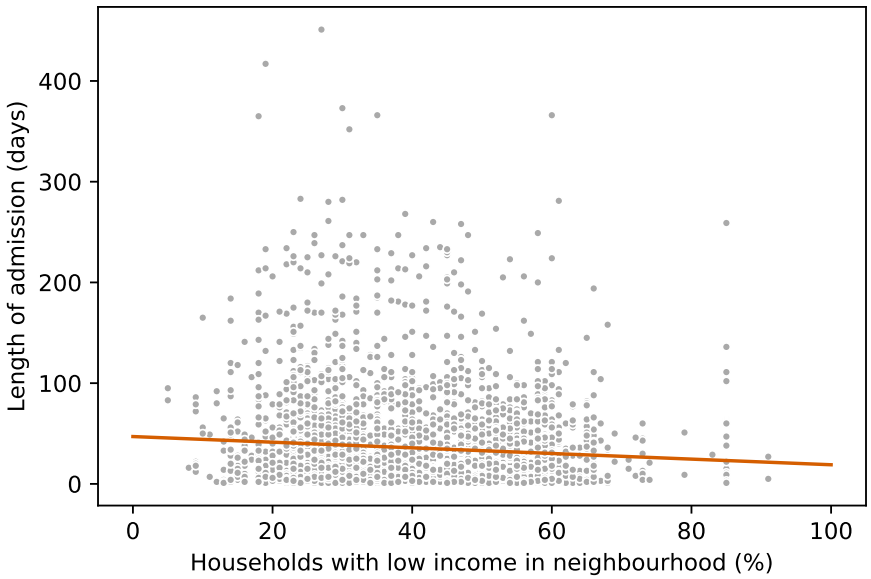


Figure 2.8: On the horizontal axis, the percentage of households with a low income in the neighborhood of a patient. On the vertical axis, the length of admission (2,238 admissions). A trend is visible towards shorter admissions for patients from a neighborhood with a high percentage of low income households, that is, low economic status.

2. Data Understanding. Some difficulties were experienced in obtaining the relevant data. The staff that supplies the data (e.g. data managers) may not be familiar with the idea of exploratory data analysis, so a clear statement of the project intentions is quite useful in order to obtain their support. Access to data was initially not very streamlined, because it was provided through several database systems and flat files that all have their own interfaces, methods, and file types. Yet eventually all necessary data were collected, and centrally stored in a uniform format.

3. General Data Preparation. The steps to be taken in this phase are generally well-known; it is however time consuming to perform them on the data. The most commonly performed tasks are well described in literature, and applying them poses no real fundamental challenges, yet they are inevitable in transforming the raw data to analyzable data. Another aspect is that data quality deficits may come to light that are related to registration issues, and resolving those deficits requires inquiring with the people that are responsible for data registration, another time consuming task.

4. Modeling and Evaluation. Focusing on one theme during each of the sessions offered a way to balance between utilizing the diversity of the data and not overwhelming participants with it. A decent introduction to the visualization tool and its different visualizations proved to be effective, demonstrated by the fact that nearly all participants were able to contribute in the exploration sessions. The specific data preparation, and visualization setup, required effort of technical staff but ultimately did not lead to significant challenges. The attitude of participants towards data analysis in this interactive manner was generally positive and open, which is further supported by the fact that during the later stages of the project the local professionals started taking initiative in asking questions that might be answered using the data outside the weekly sessions. This is in line with the ultimate goal of adopting a more data driven standard in the mental health institution.

5. Deployment. During the deployment phase, it turned out to be very helpful that health care professionals were strongly involved in the modeling and evaluation phase. The fact that they actively participated in finding the results led to a broad support in the daily work practice for implementing the outcomes.

2.4 Conclusion

The amount of data that has become available in health care enables conducting exploratory data analysis that focuses on finding new knowledge and hypotheses, instead of solving specific well-defined problems or testing existing hypotheses. We conducted a case study at the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in the Netherlands, to investigate the possibilities and limitations of this new exploratory type of data analysis. We furthermore actively involved local workforce in all steps of the process, both to guide and to strengthen the analysis, and to pave the way for an implementation. For the case study, we have proposed a specification of the CRISP-DM that we named CRISP-IDM: Cross Industry Standard Process for Interactive Data Mining. In CRISP-IDM, most importantly the modeling and evaluation phases have been contracted into one phase that requires participation of health care professionals. Furthermore, the data preparation phase has been split in a general and specific preparation phase, and an optional inferential analysis step has been added after the modeling and evaluation phase.

During the domain understanding phase, expert interviews were conducted resulting in the identification and prioritization of seven research themes in the department. In the subsequent data understanding and data preparation phases, suitable data for researching the themes was collected, and general data preparation tasks such as transforming, cleaning, integrating, and enriching the data were performed. In the modeling and evaluation phase, data visualization was used as a tool to explore the data in collaboration with the health care professionals who guided the analysis in weekly sessions. Technical staff was responsible for selecting initial datasets, performing specific data preparation tasks that were discovered during the iterative exploration, and setting up the visualizations.

A total of 19 exploratory sessions were held with 18 different health care professionals, resulting in two direct results that were implemented in cooperation with both management and workforce professionals. Firstly, domain experts judged that the Kennedy Axis V scoring method was too constant to be of clinical use, and secondly strongly varying extensiveness of nursing reports initiated a new nurse report writing protocol. Furthermore, 29 hypotheses for further research were found, pertaining to six of the seven research themes that were identified in the domain understanding phase. Of these 29 hypotheses, 24 had not been imagined during the initial expert interviews. We have demonstrated the viability of using our CRISP-IDM method to organize exploratory

and collaborative expert sessions, and for effectively finding new knowledge and hypotheses.

3 | Supporting Reuse of EHR Data in Health Care Organizations: the CARED Research Infrastructure Framework

Health care organizations have in recent years started assembling their Electronic Health Record (EHR) data in data repositories to unlock their value using data analysis techniques. There are however a number of technical, organizational, and ethical challenges that should be considered when reusing EHR data, which infrastructure technology consisting of appropriate software and hardware components can address. In a case study in the University Medical Center Utrecht in the Netherlands, we identified nine requirements of a modern technical infrastructure for reusing EHR data: (1) integrate data sources, (2) preprocess data, (3) store data, (4) support collaboration and documentation, (5) support various software and tooling packages, (6) enhance repeatability, (7) enhance privacy and security, (8) automate data process, and (9) support analysis applications. We propose the CApable Reuse of EHR Data (CARED) framework for infrastructure that addresses these requirements, which consists of five consecutive data processing layers, and a control layer that governs the data processing. We then evaluate the framework with respect to the requirements, and finally describe its successful implementation in the Department of Psychiatry of the UMCU along with three analysis cases. Our CARED research infrastructure framework can support health care organizations that aim to successfully reuse their EHR data.

This work was originally published as:
Menger, V., Spruit, M., de Bruin, J., Kelder, T., and Scheepers, F. (2019a). Supporting reuse of EHR data in healthcare organizations: The CARED research infrastructure framework. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, volume 5: HEALTHINF, pages 41–50. SCITEPRESS - Science and Technology Publications

3.1 Introduction

The digitization of our society is rapidly creating opportunities to use new resources for research in health care in the form of routinely collected large datasets (Murdoch and Detsky, 2013; Priyanka and Kulennavar, 2014). Meanwhile, recent advancements in Machine Learning and Big Data analytics enable unlocking the potential value of these datasets (Groves et al., 2013). While current research in health care is predominantly based on Randomized Controlled Trials (RCTs) and Cohort Studies (CSs), a data analytics approach on the other hand integrates real time data from various sources within a health care organization, such as structured patient records, unstructured text notes, lab measurements, financial data, and various others (Badawi et al., 2014; Friedman et al., 2015). This approach can have substantial benefits in addition to RCT and CS study designs, in terms of cost-effectiveness, sample size and reduction of selection bias (Gandomi and Haider, 2015; Raghupathi and Raghupathi, 2014). In the near future, this may allow a transition to a data driven health care, where using real time clinical data for supporting important decisions in the care process becomes the norm, and research using data analytics becomes an important driver of new insights into aetiology and treatment of disease (Murdoch and Detsky, 2013).

In this context, the various types of EHR data that have been gathered for the sake of delivering care to patients have become an important asset to health care organizations, that is nowadays typically available in a digital format (Dean et al., 2009). Health care organizations have therefore started to assemble their EHR data in data repositories in order to apply data analytics techniques to them (Lokhandwala and Rush, 2016; Obermeyer and Lee, 2017). Several challenges in management and analysis of data have subsequently emerged, not only of a technical nature (e.g. secure storage of data, data preprocessing, and data analysis), but also of an organizational nature (e.g. combining data from different sources, effective collaboration between researchers, and reproducibility of research), and an ethical nature (e.g. legal regulations and patient privacy concerns) (Hersh et al., 2013b; Meystre et al., 2017; Safran, 2014). To mitigate these challenges, a dedicated infrastructure consisting of appropriate hardware and software components is essential for gaining reliable and secure access to the data, and for sharing knowledge about the structure and meaning of data (Danciu et al., 2014; Jensen et al., 2012). Current data repositories however are often based on Data Warehouse (DWH) technology which falls short in addressing most of these challenges, leading to scattering of both data and knowledge about data within an organization

(George et al., 2015; Roski et al., 2014).

A dedicated research infrastructure for reusing EHR data improves on this situation by providing a unified data management practice for all researchers involved, ranging from data sources on the one hand to applications in clinical practice and clinical research on the other (Hersh et al., 2013a). For example, it helps reduce errors in analysis, improves possibilities for collaboration among researchers of various disciplines, and leads to more efficient use of time and resources in the long term (Pollard et al., 2016). Although the benefits of such an infrastructure are apparent, a general framework for an infrastructure to support reusing EHR data has not yet been proposed. Our study aims to provide such a framework. Because research infrastructure software packages need to interoperate with a large variety of health IT systems, databases, EHR software from different vendors, and other local standards (Hammami et al., 2014; Kasthurirathne et al., 2015), our study will address this problem on the conceptual level rather than offering a software solution. Individual health care organizations can subsequently use existing tools within their organization, supplemented with (open source) software packages to implement an infrastructure consisting of appropriate hardware and software components based on our proposed framework, that is interoperable with their current systems and practices.

In this study, we will thus identify the most important requirements for an infrastructure for reusing EHR data by means of expert interviews in the University Medical Center Utrecht (UMCU) in the Netherlands. We then translate these requirements into relevant concepts and their relations, and arrange them in a generic framework. The main merit of the framework we propose lies in providing clear concepts that need to be instantiated in a research infrastructure for reusing EHR data, thereby supporting learning health care organizations that aim to do so. We furthermore describe the implementation of our proposed infrastructure framework in the Department of Psychiatry of the UMCU, and present three specific applications that were enabled by this infrastructure.

3.1.1 Related Work

In contrast to research into reusing EHR data within an organization, there are various examples of projects that aim to integrate all types of clinical data from different institutions. For example, the CER Hub (Hazlehurst et al., 2015) which provides standardized access to the patient centric EHR across organizations, the SHARPn project (Rea et al., 2012) which enables using the EHR for secondary purposes in multiple academic centers, and EHR4CR

(Moor et al., 2015) which offers a scalable and efficient approach to interoperability between EHR systems. Additionally the eMERGE, PCORnet, and SHRINE projects provide more research into the ethical, legal, and social issues of combining data from multiple sites (Fleurence et al., 2014; McCarty et al., 2011; Weber et al., 2009). All of these projects address topics such as semantic interoperability, data quality and data integration through a Trusted Third Party (TTP), which are only indirectly relevant when reusing EHR data within an organization.

Data management practices within an organization are usually designed for dealing with data from CS and RCT studies (Krishnankutty et al., 2012), accompanied with infrastructure in the form of a Clinical Data Management System (CMDS) (Lu and Su, 2010). The data that is produced by CS and RCT studies contains measurements that are clearly defined in a study protocol, and that are often static after patient enrollment has ended. Challenges include data-entry and medical coding of data. This type of clinical data strongly differs from secondary EHR data, which is already present data that is updated live, and is often undocumented.

Research into infrastructure for reusing EHR, which is scarce in the first place, typically describes one or two requirements, and thereby only a small part of the solution that is needed. For example, they focus on the preprocessing and analysis pipeline (Peek et al., 2014), analyzing and storing large datasets (Youssef, 2014), integration of data sources (Bauer et al., 2016), or composing research datasets from secondary data (Murphy et al., 2010). Not one approach however provides a unifying data management practice, failing to provide the broad scope for infrastructure that we envision. The generic framework that our study will provide therefore has additional value for the field of clinical research data management.

3.2 Methods

3.2.1 Identification of Requirements

Requirements for an infrastructure for reusing EHR data were identified in the UMCU. The process of identifying these requirements from expert interviews is depicted in Figure 3.1. First, semi-structured interviews were conducted with seven relevant stakeholders in the UMCU to explore the requirements for an infrastructure. A board level stakeholder and psychiatrist, a nurse researcher, and data and IT experts with several backgrounds were included in the interviews, ensuring representation of all relevant stakeholders. In this

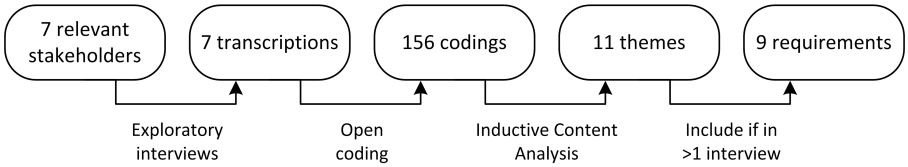


Figure 3.1: The process of identifying the nine requirements for the infrastructure.

case, semi-structured interviews are the most appropriate method for eliciting new information (Gill et al., 2008). Questions about the participants' views on current data management practices, possible improvements to these practices and their feasibility in the context of the UMCU, current issues experienced, and their ideal data management situation were asked.

The transcripts of these interviews were then processed using a grounded theory approach. Researchers first applied an open coding process to the transcripts by segmenting them, and describing each segment in a word or short sequence of words (Strauss and Corbin, 1990). This resulted in 156 codings in the seven interviews combined. To be able to identify requirements out of these codings, the open codings were then processed into broader categories by grouping codings based on their similarities and differences using Inductive Content Analysis (Elo and Kyngäs, 2008). This iterative process resulted in eleven themes that spanned multiple codes, and could not be further combined. In a final step, all themes that were mentioned by more than one interviewee were reformulated into requirements. Table 3.1 shows the roles of the interviewees within the UMCU, and the nine requirements that were identified based on the interviews, including the number of codings for each interview and theme.

3.2.2 Requirements

The nine requirements that were identified based on the expert interviews are discussed below.

1. *Integrate data sources.* In a health care organization, typically multiple data sources exist (e.g. different database systems for structured patient records, unstructured text notes, correspondence, lab values, financial administration, and genetics), that store data in their own format, and do not necessarily communicate and interoperate with each other (Coorevits et al., 2013; Nair et al., 2016). Further integration with open data sources

Table 3.1: The number of codings in each interview, distributed over the nine themes and seven interviewees. The rightmost column shows the requirement that was formulated based on the theme in the leftmost column.

Theme	**Interviewee**							**Requirement**
	Board level, Psychiatrist	Data Manager	Information Architect	Data Analyst 1	Data Analyst 2	Data Analyst 3	Nurse Researcher	
Data sources		1	3		1	1		(1) Integrate data sources
Data standardization and preparation	7	1	6	12	2	6		(2) Preprocess data
Data storage			4	7	1	9		(3) Store data
Software and tooling	2	2	12	9	1	3	1	(4) Support various software and tooling packages
Coding best practices and documentation	3	2	2	3	1	3		(5) Support collaboration and documentation
Repeatability	1		1	2			1	(6) Enhance repeatability
Privacy and security	1					1		(7) Enhance privacy and security
Data process automation	3	3	3	6		4	1	(8) Automate data process
Health care practice applications	4	1	3	6		4	2	(9) Support analysis applications

and patient-gathered data (e.g. from wearables or social media) can offer even better insights.

2. *Preprocess data.* Data preparation is a crucial step in the data analysis process that is relatively easy to model yet time consuming, especially if its steps are repeated for each separate analysis (Priest et al., 2014; Wickham, 2014). Applying preprocessing steps (e.g. tidying, standardizing, reshaping, and integrating data) in a central, collaborative manner thus saves time and effort for all researchers involved. There is on the other hand a trade-off with flexibility, and researchers should also be enabled to make their individual choices in data preparation steps where needed.

3. *Store data.* Data that is gathered from various sources and then preprocessed should be accessible in a uniform format, allowing researchers to load necessary data into their tools for analysis (Apte et al., 2011; Jensen et al., 2012).

4. *Support various software and tooling packages.* Data analysis teams are typically multidisciplinary, consisting for example of data analysts, health researchers, practitioners, and statisticians (Lokhandwala and Rush, 2016), each applying a wide range of different techniques such as classical statistics, machine learning, and data visualization (Katal et al., 2013). This leads to a variety of different software and tooling packages being used, which all need to be able to interoperate with a central infrastructure if adoption is to be achieved.

5. *Support collaboration and documentation.* Among researchers that contribute to a data analysis project, collaboration is vital for obtaining both high quality data and analysis (Cheruvelil et al., 2014; Priest et al., 2014). This is mainly achieved by documentation and code collaboration (Wilson et al., 2014), adoption of which is currently low in health care research (Murphy et al., 2012). Documenting data firstly improves shared knowledge about data, a lack of which is one of the largest barriers for performing analysis, especially in health care (Lee et al., 2015). Code collaboration secondly reduces redundancy and errors.

6. *Enhance repeatability.* Reproducible research is slowly becoming the norm in data-intensive scientific research (Peng, 2011), yet it is still not uncommon for researchers to be unable to recover data associated with their own published works (Goodman et al., 2014; Pollard et al., 2016). Data analysis in health care requires a reproducible workflow, which has

well-recognized benefits, both internally (e.g. traceability of data, better insights into data provenance), and externally (e.g. better substantiation of results, enabling reuse of methods and results for others) (Johnson et al., 2014; Wang and Hajli, 2017).

7. *Enhance privacy and security.* Health care data that are made available for research comprise sensitive data, that should be handled securely and with respect for patient privacy by design (Gil et al., 2007; Kupwade Patil and Seshadri, 2014). Security-wise, restrictions on who can access which part of research datasets help prevent data leaks and unnecessary risks of patient re-identification. Regarding privacy, de-identification techniques (e.g. pseudonymization, de-identification of free-text variables, k-anonymity measures) are needed to mitigate impact on patient privacy (Menger et al., 2018b).

8. *Automate data process.* By automating all data processing steps, up-to-date EHR data becomes available periodically, without the need to perform additional time intensive operations before analysis is started. This additionally leads to better speed to decision (Wang and Hajli, 2017) and even better model learning (Lin and Haug, 2006).

9. *Support analysis applications.* The various applications of reusing EHR data, such as decision support, dashboarding, fundamental research, data visualization, and several others (Chen et al., 2012; Gandomi and Haider, 2015), should all be supported.

3.2.3 Framework Development

A conceptual framework for a data infrastructure was designed based on the nine requirements described above. As current data repositories are commonly based on Data Warehouse technology, the DWH model of Inmon (2002) was used as a starting point. This model defines four data layers: the Data Source layer, the Staging layer, the Data Warehouse layer, and the Data Access layer. The layers in this model were iteratively refined, separated, and combined, and new layers were added in order to meet all the nine requirements. Next, these layers were integrated in a single unifying framework. This design was presented and discussed in a focus group with the stakeholders in Table 3.1, aiming to demonstrate the framework, and to evaluate it with respect to the nine requirements. A focus group is appropriate so that interaction between stakeholders is possible, so that all opinions about the framework can be explored, and so that all its potential issues are found (Gill et al., 2008). One of

the researchers facilitated the focus group, while the stakeholders were present to discuss the requirements and the framework. Comments mainly concerned the extent to which data preprocessing can be done in advance, the privacy steps that needed to be taken, and the viability of implementing infrastructure based on the framework in the UMCU. The participants attitude towards the framework was generally positive, and based on this focus group no major changes to the framework were introduced.

An infrastructure based on this framework was finally implemented in the Department of Psychiatry of the UMCU. There, an initiative to bring data driven research to the daily practice resulted in some preliminary results (Menger et al., 2016), but with no further supporting infrastructure in place, making it an ideal case for implementing the framework.

3.3 Results

3.3.1 Framework

The CApable Reuse of EHR Data (CARED) framework we designed can be seen in Figure 3.2. It consists of five data layers (from left to right) in which data is processed, and a control layer for governing the data process.

In the data processing layers, first the *extract layer* connects to various internal and external data sources, and extracts the data in their own format. The subsequent *privacy layer* performs operations that help guarantee patient privacy, such as de-identification, pseudonymization, and removal of non-consenting patients' data. Next, we propose to apply data preprocessing in two separate *general* and *specific* preparation steps. In the *general preparation layer*, the extracted and de-identified data from the several sources is transformed, and tidying steps such as standardization, reformatting, and reshaping are applied. The data remains semantically unchanged, meaning that only transformations concerning the format of data are performed. The cleaned data is stored in an analysis data repository, where it can be accessed by researchers through an Application Programmable Interface (API). In the *specific preparation layer*, advanced transformations that require domain knowledge, such as imputation, data integration, and data enrichment are applied to data from the analysis data repository. After this phase, processed data are stored in a second, context-specific data repository that is also accessible through an API. In the final application layer, data from the analysis data repository or context-specific data repository are processed further by individual researchers according to their analysis purpose.

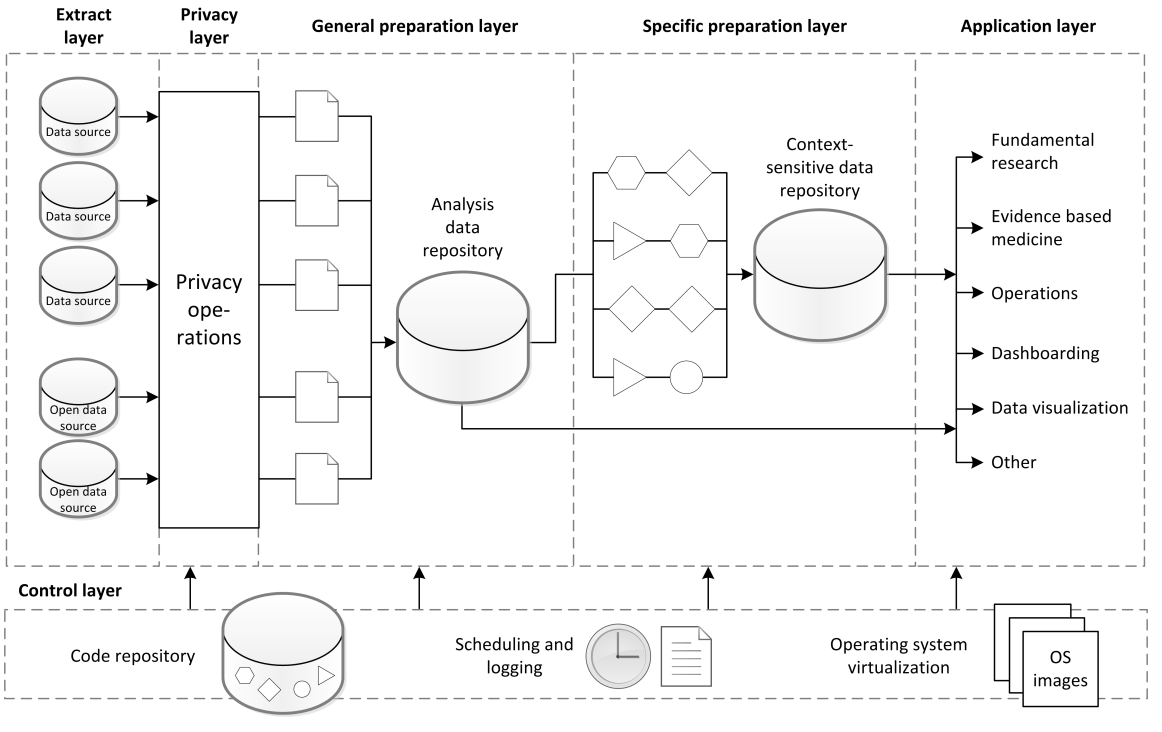


Figure 3.2: The CARED framework, consisting of five data processing layers from left to right, and an additional control layer that governs the data processing.

Additionally, the *control layer* drives the data through the data processing layers. This layer consists of three parts. Firstly, it contains a code repository where all code that transforms the data in or between layers is collaboratively written and maintained by all researchers involved. This creates a scientific workflow that traces data and code from source to application when guidelines for documenting data and code with version control are applied. Secondly, the scheduling and logging component makes sure that all clinical data is periodically extracted from the data sources and processed as specified in the code base. Reporting and notification ensure that errors in this process can be noticed and corrected. The execution of code thirdly is performed by using containerization on the operating system level. Images that provide all necessary software and libraries to execute code from the repository in a container are specified, so that the data process is not dependent on individual researchers' software.

3.3.2 Evaluation

Below, we will describe how and in which parts of the framework the identified requirements are addressed. Between square brackets, the abbreviated layer(s) in which this requirement is satisfied is written (e = extract layer, p = privacy layer, gp = general preparation layer, sp = specific preparation layer, a = application layer, c = control layer).

1. *Integrate data sources [e].* Multiple data sources are integrated in the extract layer, which allows flexibility with regard to adding or removing data sources.

2. *Preprocess data [e, gp, sp].* We propose to divide the preprocessing of data into two steps: a general data preparation step and a subsequent specific data preparation step. In the first step, operations concerning the format of data are performed, and in the second step additional operations concerning the contents of data are performed. This distinction allows researchers to choose between two datasets with a different level of preparation, balancing between flexibility and time-efficiency. In the first case, all preprocessing operations are performed by the individual researcher, but can be tailored to specific needs, while in the second case off-the-shelf preparation reduces both the effort needed to start analysis and the likelihood of errors. In both cases, the researcher needs to perform final analysis-specific preparation in order to perform the analysis.

3. *Store data [gp, sp].* Data are stored in an accessible location, in two

analysis data and context-sensitive data repositories. An important requirement for data storage is that read- and write access can be provided for relevant software packages through an API. The data storage method can be subject to organizational and technical requirements, with options ranging from a shared drive, to a database scheme (e.g. NoSQL), or distributed file systems. To ensure that previous versions of datasets are retrievable, snapshots of data can be stored using data differencing techniques.

4. *Support various software and tooling [a, c].* The infrastructure is not dependent on specific software packages. This means that both running the data through the five data layers, and performing analysis can be performed using any software package that can access data in the two repositories through the API.

5. *Support collaboration and documentation [c].* Container images, and documentation of code and data are shared in the code repository. Access to the central code repository for all researchers enables collaboration both in the data process, and in specific research applications.

6. *Enhance repeatability [e, gp, sp, a, c].* Firstly, the pipeline structure of the framework ensures that all data can be traced back to its source, providing data lineage for all eventual applications. Secondly, previous versions of data, code, and operating system containers that are all stored together create a scientific workflow, which allows repeating analysis internally. By making the combination of these three items publicly available (e.g. along with a published result), analysis additionally becomes repeatable for the entire research community.

7. *Enhance privacy and security [p].* Privacy of patients is enhanced by incorporating a separate privacy layer as one of the first data processing layers, ensuring that no datasets with identifying information proliferate in the analysis process. Security of data is enhanced by storing data in two central repositories that can be accessed by an API, preventing creation of copies of datasets on personal drives. Both the data access API and any other interfaces that allow interaction with the infrastructure can have restrictions on access for individual researchers, for example when only a subset of data is relevant for a specific analysis goal.

8. *Automate data process [c].* All code that drives the data through the processing layers is available in the code repository, and can therefore be executed in its specified container at defined time intervals by an

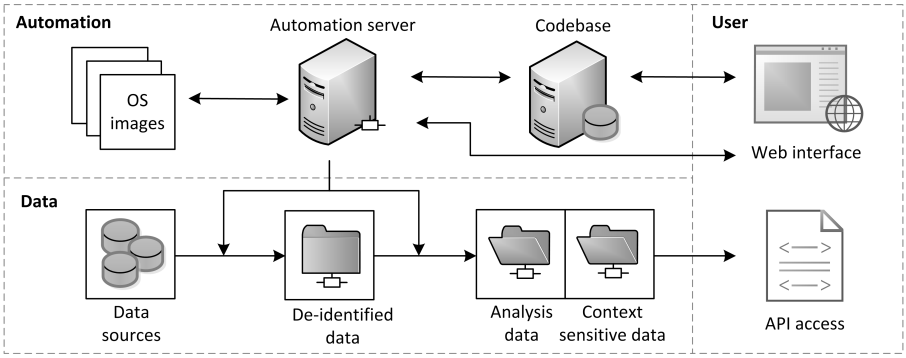


Figure 3.3: A schematic overview of infrastructure implemented in the Department of Psychiatry of the UMCU based on the CARED framework. It consists of an automation server and codebase that can be accessed through a web interface, and several data components that can be accessed through an API. Abbreviations: OS = Operating System, API = Application Programmable Interface.

automation server. This ensures that data in both data repositories are periodically updated.

9. *Support analysis applications [a].* The data in the two repositories can support a broad range of applications that is able to read data from one of the two data repositories through the API, such as data visualization, dashboards, (re)training of machine learning models, and decision support in the EHR.

3.3.3 Implementation

Based on this framework, an infrastructure was designed and implemented in the Department of Psychiatry of the UMCU in a time period of 6 months. The subsequent programming of data processing pipelines took another approximate 6 months. Software components were implemented with open source packages such as GitLab, Docker, Jenkins, Python, and R, supplemented with already present enterprise software to extract data from internal data sources. The hardware setup consists of a Linux and a Windows server, in order to allow interoperation with existing systems.

The implementation is schematically depicted in Figure 3.3. The core of the implementation is the automation server, which updates data weekly by fetch-

ing code and OS containers, and applying it to data. The data is extracted, de-identified, and stored in a network folder. Then additional preprocessing steps are applied, and another network folder stores the data in analysis data and context sensitive data repositories. After mounting the network folder, researchers can access data in these repositories through an API based on their authorization. The code base and automation server are accessible through their web interfaces.

- *Personalized antipsychotics and antidepressants prescriptions.* Patients with a psychotic or depressive disorder are often prescribed medication as part of their therapy. There are however various types of both antipsychotics and antidepressants, and choosing a drug and dose that improve an individual patients symptoms while minimizing side effects is mostly based on trial and error. During their lifetime, these patients typically switch medication and dose several times before the optimal combination is found. To smoothen this process, we developed visualizations of medication history based on information in the EHR, that is displayed to psychiatrists and patients. An overview of previous steps taken in finding an optimal prescription are comprehensively displayed, facilitating better decisions about next steps.

- *Prediction of inpatient violence incidents.* During psychiatric admission, violence from patient directed at staff or other patients can occur. This topic has been thoroughly researched in psychiatry literature, yet a data driven approach had not been applied. We integrated admission data, textual data in the first 24 hours of admission, and violence incident reports, and then applied machine learning to train a classifier that is able to assess violence risk for individual patients, outperforming trained psychiatrists and other existing violence risk assessment tools (Menger et al., 2018a). This experiment is fully repeatable, while predictions can eventually be shown in the EHR.

- *Tracking patient enrollment status.* Before a patient can start a planned admission or therapy, several administrative steps need to be taken that can take a span of multiple weeks or months. This includes for example obtaining referral documents from previous care organizations or a general practitioner, checking health insurance, and planning admission or therapy. Based on the current status of enrollment as written in text notes in the EHR, we therefore designed a status tracking system for patient enrollment. This leads to better insight into a current enrollment status, which benefits both patients and staff.

3.4 Discussion

To fully realize the potential of analyzing already existing EHR data, an infrastructure consisting of appropriate hardware and software components is needed, so that important technical, organizational, and ethical challenges of reusing EHR data are mitigated. Current data repositories in health care organizations are often still based on DWH technologies, which fall short in addressing many of these challenges. Our CARED framework, designed based on requirements that were identified in the UMCU, provides a modern and unifying approach to infrastructure for EHR data reuse. It addresses important challenges, that are too often disregarded or solved in an ad hoc manner, such as analyzing sensitive data with regard for patient privacy, repeatability of analysis, collaboration among researchers, and documentation of data and its analysis. Current research typically manages one or two of these challenges, while our research provides a framework that covers all the important aspects of reusing EHR data. We argue that adopting this framework improves quality of analysis, enhances patient privacy and data security, and aids efficient use of time, resources, and skills. By providing a generic framework, we furthermore circumvent problems of interoperability with current IT systems, improving likelihood of its adoption. Adhering to the CARED framework when designing and implementing infrastructure in a health care organization will therefore be able to improve the state of data analytics research on secondary EHR data.

Implementing an infrastructure that is based on our proposed framework in the Department of Psychiatry of the UMCU furthermore shows the feasibility of such a project. Although organizational factors caused some delays and practical difficulties, no fundamental setbacks were experienced. Additionally, we made use of existing open source software packages, leveraging knowledge and efforts from the extensive ecosystem of data analysis researchers. This is an important benefit that remains unaddressed in other software solutions. Modern data analysts are well versed in performing analysis using open source packages, typically implemented in the Python and/or R programming languages. Using such open software packages is a cost-effective measure that additionally lowers the threshold for data analysts from various domains to join the challenge of obtaining value from EHR data, which holds many promises for the future.

3.5 Conclusion

In this study, we used expert interviews to identify the most important requirements for an infrastructure for reusing EHR data, and subsequently designed the CApable Reuse of EHR Data (CARED) framework for infrastructure that addresses these challenges. The CARED framework we propose consists of five data processing layers: an extract layer, a privacy layer, two preprocessing layers, and an application layer. The framework is governed by a control layer, which consists of a code base where code and analysis is documented, a scheduler that automates the process, and containerization to make the analysis more robust and repeatable. We have elaborated upon the implementation of an infrastructure based on the proposed framework, showing its feasibility. Our study shows how an infrastructure based on the CARED framework in place will improve the quality of analysis, enable types of analysis that are otherwise not possible, and aid efficient use of time, resources, and skills.

4 | DEDUCE: A Pattern Matching Method for Automatic De-identification of Dutch Medical Text

In order to use medical text for research purposes, it is necessary to de-identify the text for legal and privacy reasons. We report on a pattern matching method to automatically de-identify medical text written in Dutch, which requires a low amount of effort to be hand tailored. First, a selection of Protected Health Information (PHI) categories is determined in cooperation with medical staff. Then, we devise a method for de-identifying all information in one of these PHI categories, that relies on lookup tables, decision rules, and fuzzy string matching. Our de-identification method DEDUCE is validated on a test corpus of 200 nursing notes and 200 treatment plans obtained from the University Medical Center Utrecht in the Netherlands, achieving a total micro-averaged precision of 0.814, a recall of 0.916 and a F_1-score of 0.862. For person names, a recall of 0.964 was achieved, while no names of patients were missed.

This work was originally published as:

Menger, V., Scheepers, F., van Wijk, L. M., and Spruit, M. (2018b). DEDUCE: A pattern matching method for automatic de-identification of dutch medical text. *Telematics and Informatics*, 35(4):727–736

4.1 Introduction

Data from EHRs are since long being used for medical research purposes, and with the increasing digitalization in the medical world even more so (Milovic, 2012; Murdoch and Detsky, 2013). Since many hospitals today have adopted an EHR system, the produced data has high potential to be used for clinical research (Koh et al., 2005; Lee et al., 2013). Both health care institutions and patients can directly benefit from the results of this kind of research that can improve diagnosis, treatment, hospital operations, and more (Jensen et al., 2012; Menger et al., 2016). The use of patient data for research however puts a strain on patient privacy, since this requires getting the data out of their health care context (Patel et al., 2015). This entails for example copying the data to different databases, where it can be accessed by data analysts (Simon et al., 2000). Technical staff such as data managers or data analysts typically do not have a treatment relation with the patient, and therefore should not be able to identify individual patients in a research dataset. Medical staff on the other hand is allowed to see patient information under medical confidentiality, but lack the technical skills to perform advanced analysis that is needed for obtaining direct clinical value from data.

Multiple ways exist to solve this problem for structured data. One rigorous approach is to remove all variables that could identify a person, such as patient names, addresses, and social security numbers from the dataset (Emam et al., 2006). A more sophisticated solution is a k-anonymity method, where the information of a patient is indistinguishable from at least $k-1$ individuals (van Toledo and Spruit, 2016). In recent years however, the more widespread availability and quality of text mining approaches have shifted attention to analyzing unstructured textual data in addition to structured data. It is becoming ever more apparent that these approaches offer substantial benefits for data driven research (Harpaz et al., 2014; Maenner et al., 2016; Patel et al., 2015), and that textual data from the medical domain holds valuable information that should not be disregarded. Therefore, if we require research datasets to be anonymous to mitigate the potential negative impact on patient privacy, we must also de-identify the medical free text variables.

From a patient perspective, protecting the private details of a disease from the public is essential in retaining the trust bond between a physician and the patient (Krishna et al., 2007). Any violation of this confidentiality can therefore have serious consequences for the relation between a health care institution and a patient. A patient may be adverse to their data being used for research when non medical staff has access to private details, and might

even consider seeking treatment elsewhere. Moreover, a potential data breach may expose private patient information to the general public. In the USA alone, between 2010 and 2013, 29,000,000 patient records were compromised (Liu et al., 2015). Clearly, such events have serious consequences for both the hospital and the patient.

From a legal perspective, on a European level the Directive 95/46/EG of the European Parliament on the protection of individuals with regard to the processing of personal data and on the free movement of such data was introduced in 1995 (European Data Protection Directive, 1995). All members of the European Union are obliged to take this directive into account, but how they implement it varies for each member. For example, in Sweden regional Ethics Committees give permission for the reuse of EHRs if the information that can identify a patient is removed (Velupillai et al., 2009). Similar measures are implemented in France in the French Data Protection Authority (Grouin et al., 2009). In the USA, the stricter Health Insurance Portability and Accountability Act (HIPAA) protects the privacy of health care data, requiring 18 different categories of identifying information ranging from person names to biometric identifiers such as fingerprints to be removed from medical data (HIPAA, 1996).

In the Netherlands no specific laws on the reuse of medical data exist, but there are general rules for dealing with personal data, that can be applied to medical data as well. Since EHR data is used for retrospective research, is not specifically collected for research purposes, and human subjects are only indirectly involved, the Medical Research Involving Human Subjects Act (WMO) usually does not need to be taken into account. Only the Agreement on Medical Treatment Act (WBGO) and the Personal Data Protection Act (WbP), which is the implementation of the European Directive 95/46/EG mentioned above, play a role in this situation. Retrospective research with medical records needs to be proposed to the medical ethics committee (METC), which verifies that the proposed research is in line with privacy legislation. An exception to this is when only anonymized data is used, which is the case if the de-identification process is executed perfectly.

For the two reasons above it is therefore important to de-identify as much of research data as possible, both to retain patient privacy, and to be able to comply with legal requirements. For de-identification of personal data, a distinction can be made between directly identifying information and indirectly identifying information. Directly identifying information, such as names, phone numbers, and citizen service numbers allow identification of a person with just that information. Indirectly identifying information, such as postal codes and birth dates, is not directly relatable to a person, but if pieces of

indirectly identifying information are combined it is easy to identify someone (Borking and Raab, 2001). In medical text data both directly and indirectly identifying information can be present, and both types of information need to be removed to successfully de-identify medical text. Although manual de-identification is possible, it is time consuming and generally prone to error, while automatic de-identification is feasible and easily scalable to large numbers of records (Deleger et al., 2013). For this reason, we choose to develop an automatic de-identification method. Our goal is to remove as much de-identifying information as possible, while ensuring the de-identified text is still human readable, so that research can still be carried out. Even strict de-identification methods still retain good readability of the remaining text (Meystre et al., 2014). We therefore strive to balance towards developing a method with a high recall while also maintaining a good precision.

Since there are differences in legislation that exists on a national level, and because of language-specific problems that occur in the different the types of identifying information, it is clear that a separate de-identification method must be developed for each language (Grouin et al., 2009). Although many research into the de-identification problem has been performed in English, a reliable method for de-identifying Dutch medical text has yet to be developed. For the English language, most notably Neamatullah et al. (2008) obtained a recall of 0.967 using their pattern matching method that was developed on a test corpus of 1,836 nursing notes. Uzuner et al. (2008) managed to achieve a 0.97 F_1-score on medical discharge summaries, based on a machine learning approach. A hybrid approach was developed by Ferrández et al. (2012a), achieving a 0.922 recall by combining both pattern matching and machine learning techniques. Many more approaches in English exist (e.g. Douglass et al., 2005; Fenz et al., 2014; Friedlin and McDonald, 2008).

Apart from text-processing the English language, Velupillai et al. (2009) attempted to port the Neamatullah et al. (2008) algorithm to Swedish, but in their own words with "poor results". The average score over all de-identified categories was a 0.65 F_1-measure. Over a decade ago already, Ruch et al. (2000), successfully managed to de-identify discharge letters written in French with a recall of 0.98 using their MEDTAG framework for semantic tagging. For notes that are written in Korean and English, Shin et al. (2015) developed a method based on regular expressions that achieved a 0.963 recall on several categories combined. In Dutch, Scheurwegs et al. (2013) managed a recall of 0.89 on a previously unseen dataset with a machine learning approach, achieving reasonable success using limited training data.

As can be seen, the two most common methods to de-identify medical text are pattern matching based or machine learning based (Meystre et al., 2014).

In the former, lookup tables and decision rules are used to determine what parts of the text contain identifying information. In the latter approach, machine learning techniques are used to classify each piece of text. A third option is a hybrid approach, combining the two. In literature, it is not immediately clear whether pattern matching or machine learning based methods perform better, although hybrid approaches generally outperform other approaches. A clear downside of the machine learning approach (and therefore also the hybrid approach), however, is the need for a large annotated training corpus, which requires extensive manual labor, and is thus expensive to obtain (Neamatullah et al., 2008). For this reason, such a corpus is currently unavailable in Dutch. As Ferrández et al. (2012b) furthermore shows, pattern matching based methods may even achieve better recall on unseen data than machine learning approaches. For these reasons, we opt to develop a pattern matching based de-identification method.

The development of our DE-identification method for DUtch mediCal tExt, that we name DEDUCE, will be structured along the Design Science Research Process (DSRP) methodology (Peffers et al., 2007). The first two phases comprise the identification of the de-identification problem and the objectives of a solid de-identification method, largely described above. In Section 4.2, the phases 3 and 4 concerning the design, development, and demonstration of our de-identification method DEDUCE will be described. We will elaborate on the selection of relevant Protected Health Information (PHI) categories as well as the data that we used to develop and test our method, and in detail describe how all PHI categories are de-identified. In Section 4.3 finally, we will describe how the method is evaluated and deployed (phases 5 and 6).

4.2 Method

4.2.1 Development Corpus

The de-identification method is developed and tested on data from the Department of Psychiatry the University Medical Center Utrecht (UMCU) in the Netherlands. Specifically, nurse notes and treatment plans in the period January 2012–December 2015 are made available. Nurse notes are written by nurses about all inpatients during each of the three daily shifts, and include information about the current well-being and activities of the patients. The treatment plan is typically written at the start of a treatment, and describes the activities that will take place in the context of an admission or outpatient treatment, such as therapies or medication prescription. While the nurse notes

Table 4.1: Descriptive statistics about the two data sources used to develop and test the de-identification method.

	Nurse notes	**Treatment plan**
No. records	113,553	4,012
No. patients	1,452	2,025
Avg. no. words	100	658
Type of patient	Inpatient	Inpatient and outpatient
Written by	Nurse	Physician
Type	Free-text	Free-text

tend to focus on the daily routine, and thus mention names of patients and treatment staff frequently, a treatment plan has a stronger focus on the long term treatment, and therefore more often mentions locations and other health care institutions where a patient has previously been treated or might be referred after treatment. Using both these data sources ensures a diverse corpus is used for developing and testing the de-identification method, and thus improves the generalizability to other medical text written in Dutch. Some more descriptive statistics about the two data sources can be found in Table 4.1.

To the records of both the nurse notes and the treatment plans the first names and last names of a patient as described in the EHR system are added. The resulting data set comprises a total of 113,553 nurse notes and 4,012 treatment plans. From both data sources 1,000 records are sampled at random, the 2,000 resulting records will comprise the development corpus. To ensure our de-identification method is not biased, the method will be developed on this development corpus without using any other records, and later be validated on a disjoint test corpus.

4.2.2 PHI Selection

Since no clear guideline on which PHIs to remove exists in the Netherlands, we decided to base our method on the strict HIPAA guidelines that are implemented in the USA. In the HIPAA guidelines, 18 patient characteristics are identified: (1) names; (2) all geographic subdivisions smaller than a state; (3) all elements of dates except years and all ages above 89; (4) telephone numbers; (5) fax numbers; (6) electronic mail addresses; (7) social security numbers; (8) medical record numbers; (9) health-plan beneficiary numbers; (10) account numbers; (11) certificate and license numbers; (12) vehicle identifiers and serial numbers, including license plate numbers; (13) medical device iden-

tifiers and serial numbers; (14) URLs; (15) Internet Protocol (IP) addresses; (16) biometric identifiers including fingerprints and voiceprints; (17) full-face photographic images and any comparable images; (18) any other unique identifying number, characteristic, or code. These guidelines do not only apply to the patient, but also to relatives, household members and hospital staff.

During initial exploration of the de-identification problem, by means of manual inspection of the data, and conversation with hospital staff, not all of the 18 categories were found to occur in our medical dataset. To come up with a subset of PHIs to de-identify in our method, four health care professionals that work with the text data in our corpus on a regular basis were asked to score each PHI category as occurring "never", "sometimes", or "regularly" in either the nurse notes or the treatment plan. In these surveys, seven of the 18 PHIs were indicated to be present "sometimes" or "regularly" by at least one of the four health care professionals. During a qualitative follow up, participants indicated that in addition to these seven PHIs, names of institutions where patients are treated can be mentioned in the data as well, possibly revealing something about the identity of a patient. Our de-identification method therefore focuses on the following eight PHI categories—other categories fall outside the scope of this paper.

1. Person names, including initials
2. Geographical locations smaller than a country
3. Names of institutions that are related to patient treatment
4. Dates
5. Ages
6. Patient numbers
7. Telephone numbers
8. E-mail addresses and URLs

In the next subsections, the annotation each of these PHI categories will be elaborated upon. An overview of the de-identification method DEDUCE is depicted in Figure 4.1.

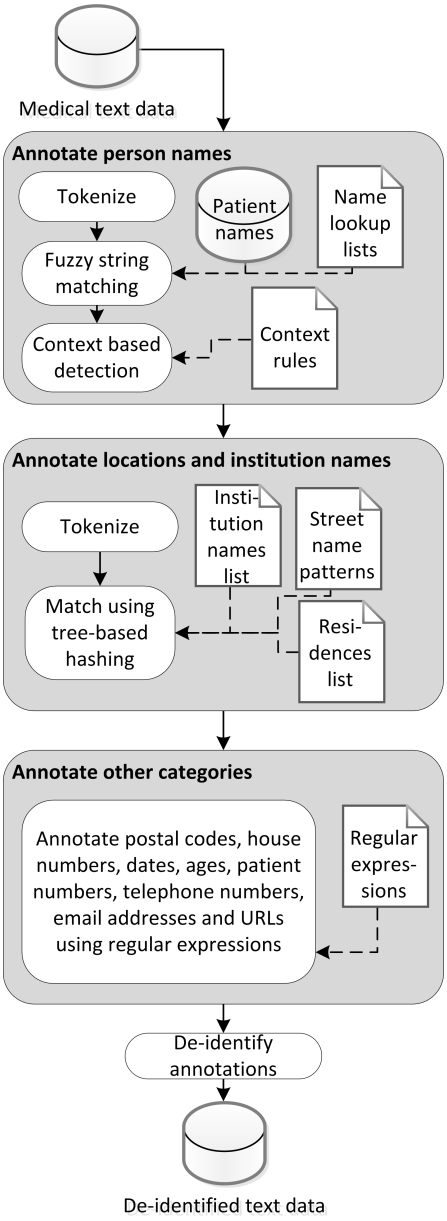


Figure 4.1: An overview of the de-identification method DEDUCE.

4.2.3 Person Names

Person names are the most common PHI in our dataset, including both first names and last names of patients, family members, and hospital employees. Only person names are included in this category, names of for example medication are not annotated.

As a preprocessing step, all non-ascii characters are mapped to their ascii counterparts, and all text is tokenized using a custom tokenizer that segregates based on sequences of alpha characters separated by non-alpha characters. The assumption behind tokenizing this way is that person names are always sequences alpha characters with the exception of prepositions that can occur in Dutch names (such as 'van der', which can be abbreviated as 'v/d' or 'v.d.'). Therefore, a list of prepositions is obtained, which will be regarded as one token by the tokenizer.

Non-context-based

The most logical first step in detecting names is using the first names and last names that are available in the EHR system. It is common for a patient to have multiple first names and/or multiple last names. For matching names to tokens, we use fuzzy string matching with a Damerau-Levenshtein (DL) distance of at most 1, which means at most one character insertion, deletion, or swap is allowed to transform the string into one of the names. By setting a threshold of 1, essentially all simple misspellings of a name are captured. Although in some rare cases two misspellings occur in a single name, increasing the threshold to 2 greatly increases the number of false positives.

Furthermore, for names with three or less characters we require an exact match to occur, since otherwise short Dutch names (such as 'Jan') are frequently matched with common Dutch short words (such as 'aan' or 'van'). For matching the names of patients in the EHR system we use three rules:

N1. Tokens that can be matched with a first name are annotated as `<PATIENT FIRST NAME>`.

N2. Sequences of tokens that can be matched with a sequence of last names (i.e. obtain a fuzzy match for each token) are annotated as `<PATIENT LAST NAME>`.

N3. A token that is equal to the first character of one of the first names is annotated as `<PATIENT INITIAL>`.

Most patient names can be annotated using the above three rules, however more challenging are names of co-patients, hospital staff, or family members. To be able to annotate these names, we obtain the following lookup lists:

- For first names, we obtain a list of the most common 10,000 Dutch first names, and a list of the most common 10,000 Dutch last names. The number of names that are needed to successfully de-identify text varies in literature from 1,8 million (Thomas et al., 2002) to 10,000 (Velupillai et al., 2009), and although it is unclear where the optimum is, shorter lists appear to perform better. The two lists are obtained from Netwerk Naamkunde of the Meertens Institute (Meertens Instituut, 2016).
- In Dutch, it is not uncommon to have a preposition between the first and last name such as 'van', 'van der', or 'in de'. A list of all prepositions in use in the Dutch language is obtained from the internet.
- A list of prefixes, such as 'dhr' (sir) or 'mw' (madam) is obtained, since they contain valuable clues about a possible presence of a name. During the development of the method, other prefixes such as 'pt' (short for patient) and 'vpk' (short for nurse) are manually added to this list.
- In order to prevent over-annotation, a whitelist of words that should not be annotated as names is compiled, consisting of the 1,000 most common words in Dutch, a list of stop words, and a list of medical terms.

For matching tokens with values on lookup lists, exact string matching is used—this includes capitalization. Empirically, fuzzy string matching as described above results in a huge amount of false positives. Using the lookup lists, the following rule for annotating names is added:

N4. Tokens that are on the list of first names or on the list of last names are respectively annotated as `<UNKNOWN FIRST NAME>` or `<UNKNOWN LAST NAME>`.

Context Based

Based on the lists of prefixes and prepositions, clues about possible names enable annotating mentions of unknown persons.

N5. If a token is on the list of prefixes, and the next token starts with a capital, the token is annotated as `<UNKNOWN PERSON>`. This includes for example occurrences like 'Mr. Smith' or 'patient Jones', whose relation to the patient cannot be automatically determined.

N6. If a token is on the list of prepositions, and the next token starts with a capital, the tokens are annotated as `<UNKNOWN LAST NAME>`.

Although a reasonable share of the names present in the text can be annotated using the rules above, a lot of names are still under-annotated, meaning that the annotation needs to be extended to the next or previous token(s) to completely annotate the name. For this, the following rules are added:

N7. If a token is a single capital letter, and the next token is annotated as either an initial or a last name, the token is annotated as an initial. This rule captures initials in front of last names, or multiple consecutive initials.

N8. For tokens that are on the list of prepositions, the context is checked to contain any annotated names. For example, if the previous token is annotated as first name or an initial, and the next token starts with a capital, it is annotated as a last name. Similarly, the token before a preposition can be annotated as a first name or initial, if the next token is annotated as a last name.

N9. If a token is annotated as an initial, first name or last name, and the next token starts with a capital, it is annotated as a last name. Although this rule produces some false positives as well, it captures an important amount of names that were not annotated because they do not occur on the lists of common last names.

N10. Finally, a token that is preceded by an annotated name and the token 'en' (and) is annotated as a name. During development, sentences like "Patient A spent its day with Patients B and C" were observed, where especially the C was commonly not annotated.

Naturally, the rules above only apply to names that were not filtered using the lists of first and last names because they are not so common in the Dutch language. In all cases, the newly annotated tokens are only annotated if they are not on the whitelist. The difference between a patient name and an unknown name is retained, by tagging the new annotation with the appropriate name based on in which context it was annotated.

Based on these ten rules (N1–N10), a very large part of the development corpus is de-identified of names: a manual scanning of a random subset of 500 pieces of text is found to contain four false negatives. Exact results will be elaborated upon in the Section 4.3.

4.2.4 Geographical Locations

Locations that are present in the data may not immediately identify a patient, but combined with other data might reveal something about the identity of a patient. For this reason, all addresses in our dataset are annotated. In Dutch, an address contains a street name, a house number with possible suffix, a postal code and a place of residence.

For annotating places of residence a list of places of residence is compiled by combining a list of all Dutch cities and villages, and a list of major European cities. The list contains a total of 2,647 places of residence. Street names can be annotated using a regular expression that matches Dutch suffixes such as 'straat' (street), 'laan' (lane), or 'plein' (square).

Since removing all values on the lists from each piece of text is computationally not feasible, for this a trie-based hashing technique is used. Initially, all locations are tokenized in the same way the medical text is tokenized, after which these sequences of tokens are stored in a trie. Then, locations are detected by iterating over all tokens in the text, and by efficiently matching the longest possible sequence of tokens in the trie using hashing. With this method, a reasonable runtime for the de-identification of locations can be obtained.

Dutch postal codes adhere to a clear format: four digits followed by two characters, e.g. 1234AB. Some minor variations on this format are observed, such as 1234ab or 1234 AB, all these variations of the postal codes are matched using regular expressions.

House numbers and possible additions can also easily be matched using regular expressions, when they are preceded by a street name that is already annotated in the previous step.

Additionally, the development corpus is observed to contain at least one occurrence of a mailbox number, the format for this is the word 'postbus' (mailbox) followed by five digits—again this can be matched using regular expressions.

Occurrences in one of the above categories are annotated with `<LOCATION>`.

4.2.5 Names of Institutions

It is possible for patients to be treated in multiple care institutions consecutively or even in parallel, especially in psychiatric care this is not uncommon. For this reason, names of care institutions where a patient is treated is regarded as indirectly identifying information that we will annotate in our dataset.

For annotating names of institutions, we combined institution names from the following sources:

- Internal data about the most common institutions where patients are also treated
- Psychiatric care institutions in the region of the UMCU
- Large psychiatric care institutions in the Netherlands
- Institutions that were identified in the text during development of the method

Choosing these lists clearly limits annotating institution names to our specific dataset, and not to Dutch medical text in general. Obtaining a complete list of all health care institutions in the Netherlands however proved to be impossible. Other users of the de-identification method however can easily compile lists that suit their data, thereby enabling the possibility of hand-tailoring the algorithm to the users' specific needs.

Although most institutions' names that are mentioned in the text are on the list, institutions are not always referred to by their official name in colloquial language. To mitigate this problem, some preparations on the list are performed:

- For institution names that contain articles or prepositions in their name (e.g. 'De Hoogstraat' which translates to 'The Hoogstraat'), the institution name without the preposition (i.e. simply 'Hoogstraat') is also added to the list.
- For institutions with three or more words in their name, the abbreviation of the institution name is also added (i.e. 'University Medical Center Utrecht' can be abbreviated 'UMCU').
- Some common abbreviations of words are also substituted and added to the list, such as 'zkh' for 'ziekenhuis' (hospital).

Our final list of institution names contains 742 values. Again, to keep things computationally efficient, our trie-based hashing method is used. All institutions that are found are annotated `<INSTITUTION>`.

4.2.6 Dates and Ages

A date could for example constitute a date of birth or date of admission, which may identify a patient. As in the HIPAA guidelines, only combinations of days and months are regarded as potentially identifying information; years do not need to be de-identified. In Dutch, dates usually follow the day-month-year format. Using the following two patterns, dates can be annotated `<DATE>`:

- A number between 0 and 31, followed by a number 1–12, possibly followed by a two or four digit number. Between the number groups, white spaces, slashes, and dots are allowed.

- A number between 0 and 31, followed by the name of a month or an abbreviation (e.g. 'Jan' for 'January')

The HIPAA guidelines state that ages over 85 should be removed from medical text, we however opt to remove all ages from the dataset. In combination with other information, the age of a patient may reveal the patient's identity. Moreover, from the EHR the year of birth can easily be obtained for selections of patients if needed. Using simple patterns such as a number followed by "year" or "year old", ages can be detected and annotated `<AGE>`.

4.2.7 Patient Numbers

Patient numbers cannot directly identify information about a patient, but may allow connecting other data sources to the text data in which the patient number is found. Unfortunately, no other structure in a patient number is found than the fact that is a 7-digit number—we therefore match all 7-digit numbers, and tag these as `<PATIENT NUMBER>`. Although this strict rule results in false positives, no other less rigorous matching of patient numbers is possible.

4.2.8 Telephone Numbers, Email Addresses and URLs

Lastly, telephone numbers, email addresses, and URLs are annotated because they can clearly directly identify a patient. For all three, an abundance of regular expressions exist on the web, that after some minor tweaks generally annotate these three PHI types well. Telephone numbers are tagged `<TELEPHONE NUMBER>`, both email addresses and URLs are tagged as `<URL>`.

4.2.9 De-identifying the Annotations

After annotating the PHIs, de-identification can take place. Although it is possible to simply remove all annotated PHIs from the text, or to replace them with a default string, this reduces the legibility of the text, and is thus not preferable. We therefore choose the following method to de-identify the annotations:

- First, all adjacent annotations are merged into a single annotation if the annotation type matches. This ensures that no tags like `<NAME><NAME>` occur in the text.
- For patient names, all occurrences of `<PATIENT FIRST NAME>`, `<PATIENT LAST NAME>`, and `<PATIENT INITIAL>` are replaced by simply `<PATIENT>`.
- For all other tags, the tag is replaced by the name of the tag with a number that uniquely identifies the occurrences of the same value. For example, all occurrences of the name of a specific nurse within one piece of text are replaced with `<PERSON-1>`, while all occurrences of the name of a co-patient in that text are replaced with `<PERSON-2>`, etc. For person names, geographical locations, and names of institutions, fuzzy string matching is used to make sure misspellings or slightly different spellings are matched to the same tag. For all other PHI categories, exact string matching is used.

A Python implementation of the DEDUCE de-identification method is made publicly available in a GitHub repository.[1]

4.3 Results and Discussion

The final de-identification method is validated on a test corpus that consists of a random sample of 200 nurse notes and 200 treatment plans. The test corpus is fully disjoint from the development corpus, in other words, no text that is used to develop the method can be selected in the test corpus. The test corpus is annotated using our de-identification method, after which the annotations are validated by a human rater. This is done by visually presenting a piece of text with annotations marked in different colors for each PHI category. The rater then for each category specifies the number of false positives (i.e. pieces of text that were erroneously marked PHI) and false negatives (i.e. pieces of text

[1] `https://github.com/vmenger/deduce/`

Table 4.2: For each PHI category, the results for de-identifying the test corpus ($n = 400$). The totals displayed are based on micro-averaging over the PHI categories.

PHI	**True Positives**	**Precision**	**Recall**	F_1**-score**
Person names and initials	270	0.742	0.964	0.839
Geographical locations	13	1	0.867	0.929
Names of institutions	102	0.990	0.756	0.857
Dates	99	0.780	0.980	0.868
Ages	50	0.980	0.980	0.980
Patient numbers	5	1	1	1
Telephone numbers	3	1	0.600	0.750
E-mail addresses/URLs	0	–	–	–
Total	542	0.814	0.916	0.862

that are not marked as PHI while they contain a PHI). If a PHI is correctly annotated, but in the wrong category, a false positive is also registered.

Furthermore, since all pieces of text are written about a single patient, it regularly occurs that the same name of a patient is written multiple times, and is thus annotated multiple times. Counting all of these occurrences as separate true positives would strongly overestimate the performance of the method, and therefore only the number of unique annotations in each category is counted. Consequently, false positives and false negatives are also only uniquely counted in one piece of text.

To verify that one human rater can accurately validate the annotations, a random selection of 150 pieces of text from the test corpus were validated by a second rater. Although some minor differences occurred, no major false negatives were overlooked, and all precision, recall, and F_1-scores matched to within a small tolerable margin of error.

The results of the validation on the test corpus can be seen in Table 4.2.

From Table 4.2 it can be seen that the de-identification method achieves generally good results, with a micro-averaged precision of 0.814, a recall of 0.916, and a F_1-score of 0.862. The recall of the method shows how likely it is for a PHI to be missed, and thus how likely it is for re-identification to occur, while the precision measures the amount of false positives, thereby estimating the amount of information loss that is a result of applying the method. These results are in line with research in other languages. More importantly, for patient names, which is the most directly identifying PHI, a good recall of

0.964 was achieved, while no patient names were missed. These numbers combined show that automatic de-identification of medical text written in Dutch is possible to a high degree using our method DEDUCE.

A total of 10 occurrences of person names were missed by the method. In one case, this was the name of a school teacher which was misspelled. In other cases, names of hospital staff that were not capitalized, or abbreviated in an uncommon way (such as concatenating initials with a last name) were not properly de-identified. One occurrence of a first name, and four occurrences of last names were missed, other occurrences concerned only initials. No person names of patients were missed by the method. For locations, two misspellings of cities were missed. The most tricky PHI to correctly annotate are names of institutions where patients are treated, with a recall of 0.756. In many cases, different spellings, abbreviations, or colloquial variants of the names of institutions made annotating difficult, and in some cases the name of an institution was not on one of the lookup lists. For telephone numbers, in one piece of text the last two digits of two phone numbers in a format from a different country were not properly included in the annotation, resulting in two false negatives. Although the recall score of 0.6 seems low, this can solely be attributed to these two phone numbers, where the text after de-identification still does not reveal any identifying information other than the two last digits of the phone numbers.

For person names, both over-annotation as well as some annotation of text that are no names takes place. In the first case too much text is annotated, this mostly concerns person names followed by a word with a capital letter, in the latter case person names that are also Dutch words are annotated. For dates, some false positives occurred when the numerical part medication dosages (such as 2.5 milligram) were annotated as a date. This can relatively easily be detected in an improved version of the method, but is not included in the validation. In other categories, very few false positives occurred. Although some email addresses and URLs were present in the development corpus, none were present in the test corpus, and thus no score is added for this category.

Finally, it must be addressed that for the geographical locations, patient numbers, and telephone numbers, relatively few occurrences were found in the text. Although the recall shows that most of these occurrences were found by the algorithm, the small sample size affects the reliability of the results in these categories. This could be a topic of further research, using a dataset that is more rich in these categories.

4.3.1 Generalizability to Other Dutch Medical Text

Our method strives to be applicable to Dutch medical text in general, some fitting to the dataset that we used to develop and test the method is however inevitable.

De-identification of person names should generalize well to data of other institutions, because it relies on national lists of popular names and generic rules. It must be noted that this part of the de-identification process partly relies on the patient names that are registered in the EHR system. In case this data is not available, performance of the method may decrease. For institution names, de-identification heavily relies upon the available list of treatment institution names. Without a list that is specific to a health care institution, performance will likely not be very good—so for this PHI category obtaining a good list of institution names is essential. For the de-identification of geographical locations, dates, ages, telephone numbers, email addresses and URLs, generalizability is expected to be good, since very little information specific to our dataset was used. Patient numbers finally will most likely not be easily detected in other datasets, since they may not follow the same 7-digit format. This part of the method can however easily be hand-tailored.

To conclude, most of the method has been designed to be simple and generic, and to thusly be more generalizable to other Dutch medical text. This is also supported by for example Ferrández et al. (2012b), who showed that pattern matching based methods generally obtained a good recall on unseen datasets. The method is generic with respect to finding PHIs in all categories, except names of institutions and patient numbers, which can be hand tailored to a specific data set with relative ease.

4.4 Conclusion

There is no doubt that medical text data holds great potential for research, and that it can be a great asset in creating direct clinical value by applying data analysis techniques. Using text data however puts a strain on patient privacy: in many cases, identifiable information about patients or patient family is mentioned in text data, which should not be accessible by data analysts. Even more serious is the risk of a potential data breach, in which private details of a treatment could become known to the public.

To mitigate these drawbacks of using medical text data for research, we set out to develop a de-identification method, that we name DEDUCE: DE-identification method for DUtch mediCal tExt.

In cooperation with local medical staff, we decided to focus on (1) person names, including initials; (2) geographical locations smaller than a country; (3) names of institutions that are related to patient treatment; (4) dates; (5) ages; (6) patient numbers; (7) telephone numbers and (8) e-mail addresses and URLs. To develop and test the de-identification method, treatment plans and nurse notes from the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in the Netherlands were used to create a development corpus ($n = 2,000$) and a disjoint test corpus ($n = 400$).

Our de-identification method for person names relies on the names of patients that are known in the EHR system, lookup lists of Dutch first and last names, and subsequently applies fuzzy string matching and context based rules to annotating the names. For geographical locations and names of institutions, a trie-based hashing method was used to efficiently annotate all occurrences on lookup lists. For the other categories, regular expressions were employed to match all PHI occurrences. After annotating all PHIs, de-identification took place by replacing all annotations with appropriate de-identified strings.

Validation of our method DEDUCE on the test corpus shows generally good results, with a total micro-averaged precision of 0.814, a recall of 0.916, and a F_1-score of 0.862. Another notably good result of the method is a recall of 0.961 for names, only missing names of treatment staff, and no single patient name. Although our method is to some extent fitted to our specific dataset, we believe it will be applicable to medical text written in Dutch in general, after some simple extra preparation to hand tailor the algorithm. The validation furthermore shows that de-identification of medical text written in Dutch in an automated manner using our method DEDUCE is possible with good results.

5 | Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text

Machine learning techniques are increasingly being applied to clinical text that is already captured in the Electronic Health Record (EHR) for the sake of delivering quality care. Applications for example include predicting patient outcomes, assessing risks, or performing diagnosis. In the past, good results have been obtained using classical techniques, such as bag-of-words features, in combination with statistical models. Recently however Deep Learning techniques, such as Word Embeddings and Recurrent Neural Networks, have shown to possibly have even greater potential. In this work, we apply several Deep Learning and classical machine learning techniques to the task of predicting violence incidents during psychiatric admission using clinical text that is already registered at the start of admission. For this purpose, we use a novel and previously unexplored dataset from the Department of Psychiatry of the University Medical Center Utrecht in The Netherlands. Results show that predicting violence incidents with state-of-the-art performance is possible, and that using Deep Learning techniques provides a relatively small but consistent improvement in performance. We finally discuss the potential implication of our findings for the psychiatric practice.

This work was originally published as:

Menger, V., Scheepers, F., and Spruit, M. (2018a). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6):981

5.1 Introduction

A majority of health care providers currently digitally stores data that has been captured for the sake of delivering care in an EHR (Adler-Milstein et al., 2015b; Peters, 2017). Subsequently, health care providers have started exploring these historical datasets to improve the quality of their care (Menger et al., 2016; Priyanka and Kulennavar, 2014). Applying machine learning techniques to the various data that are gathered can, for instance, offer new insights into the etiology of disease, provide decision support to clinical professionals in the care process, aid in performing diagnosis, and improve the operations of a health care institution (Bates et al., 2014; Lee and Yoon, 2017; Murdoch and Detsky, 2013; Whitson, 2013).

The structured data in a patient record (e.g. diagnosis, medication, lab measurements) are relatively straightforward to analyze using well-known and well-researched statistical methods, in practice however, a lot of information in EHR is captured in an unstructured free text form that is more difficult to analyze (Chapman et al., 2011; Ford et al., 2017). Despite this difficulty, the merits of utilizing clinical text for research purposes are currently being discovered in many areas of research, such as adverse event detection, phenotyping, and predictive analysis (e.g. Chen et al., 2018; Garla et al., 2013; Pestian et al., 2007; Sarker and Gonzalez, 2015). These approaches use well established methods for classification of text like bag-of-words and n-grams for representing text, and Naive Bayes and Support Vector Machine models for classifying text.

Although good results have been obtained with these approaches, novel Deep Learning techniques, such as Word Embeddings and Recurrent Neural Networks, have emerged recently, challenging the superiority of these classical approaches. Recent advances have subsequently enabled applying Deep Learning techniques to Natural Language Processing (NLP) problems. Most notably, the introduction of the word2vec (Mikolov et al., 2013) and paragraph2vec (Le and Mikolov, 2014) algorithms for learning representations of text have improved state-of-the-art results in several NLP tasks (Goldberg, 2016; LeCun et al., 2015). These Deep Learning techniques are currently also being applied to clinical text, for example by Suresh et al. (2017) who included clinical text among other data types in predicting the effect of clinical interventions for Intensive Care Unit patients. Miotto et al. (2016) introduced Deep Patient, a dense vector representation of a patient through time that is partially based on clinical text, based on which good predictions of developing several diseases can be made. Other approaches have focused on extracting

medical concepts (Lv et al., 2016), Named Entity Recognition (Wu et al., 2015), or de-identification of medical text (Yadav et al., 2016). These approaches show that Deep Learning techniques applied to clinical text can yield state-of-the-art results in several cases. However, whether this generalizes to other clinical datasets, for example in different medical domains or in datasets with different sample sizes remains unclear.

In this work, we present a new case, comparing Deep Learning and classical machine learning techniques applied to classification of clinical text. We do so by performing an experimental evaluation of several techniques for representation and subsequent classification of text, applied to a novel and previously unused dataset from the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in The Netherlands. The task of this classification problem is to predict which patients will show violent behavior during their admission, based on clinical texts that are available at the start of their admission. Assessment of violence risk is a problem that causes a high burden for both patient and hospital staff. It has been described well in psychiatry literature, but has never been approached by applying machine learning techniques to EHR data.

5.1.1 Related Work

Applications of Deep Learning methods to the EHR are already numerous, both using structured data such as medication prescriptions, diagnosis and billing codes, and lab measurements, as well as using unstructured medical images. They are used for instance to perform information extraction, representation learning, prediction, phenotyping, and de-identification (Shickel et al., 2018). Gulshan et al. (2016) applied a deep Convolutional Neural Network to detect diabetic retinopathy in images of the retina, showing that the judgment of licensed ophthalmologists can be matched, while Esteva et al. (2017) used the same type of neural network to classify types of skin cancer with performance that matches that of board-certified dermatologists. Another approach by Lipton et al. (2015) using Recurrent Neural Networks managed to diagnose the most common conditions in Intensive Care Unit patients better than several baselines using irregularly performed measurements. Hammerla et al. (2016) applied Restricted Boltzmann Machines to data from wearable devices, outperforming other methods of monitoring the state of patients with Parkinson Disease. In addition, other different approaches applied to various types of structured clinical measurements exist (e.g. Jacobson and Dalianis, 2016; Li et al., 2014; Liang et al., 2014; Nickerson et al., 2016).

Violence from patients directed at staff or other patients is seen in almost

any psychiatric treatment facility. Iozzino et al. (2015) reports the prevalence of a violence incident happening during admission in 35 different facilities, which lies between 2% and 44%, averaging 17% over these different sites. Individual patient factors that are associated with violence risk, such as a history of substance abuse and a history of violent behavior are well described (e.g. Amore et al., 2008; McDermott et al., 2008; Papadopoulos et al., 2012; Pfeffer et al., 1985; Reynolds et al., 2013). Meta analyses, however, reveal that only a small amount of these factors is found to be robust when comparing the results of various studies on different populations (Dack et al., 2013; Steinert, 2002). The task of predicting the occurrence of violence incidents has been shown to be even more challenging, especially when no structured instruments are used (Ægisdóttir et al., 2006). Psychiatrists in training for example do not perform much better than random at this task with an Area Under Curve (AUC) of the Receiver Operator Curve of 0.52, while using a structured instrument improves the performance (AUC = 0.67) to the level of a trained psychiatrist (AUC = 0.71) (Teo et al., 2012). A substantial proportion of health care professionals therefore makes use of risk assessment instruments (Higgins et al., 2005), of which the Violence Risk Appraisal Guide (Quinsey et al., 1998), Structured Assessment of Violence Risk in Youth (Borum et al., 2005), and Historical Clinical Risk Management-20 (Douglas et al., 2014) are most commonly used (Fazel et al., 2012). A meta study by Singh et al. (2011) reports that the median predictive performance of these three instruments falls in a relatively narrow range between 0.70 and 0.74 AUC, showing that prediction of violence incidents is possible with moderately good results. There is however also a large variation in performance of these instruments over different sites (Yang et al., 2010). It seems that the heterogeneity of psychiatric patient populations inhibits straightforward generalization of measuring instruments' predictive performance to other treatment facilities. This has caused serious discussion, and sometimes even skepticism on the usability of these instruments in practice (Campbell et al., 2009; Maden, 2003). Given that these instruments are furthermore considered to be time-consuming, and thus expensive (Viljoen et al., 2010), predicting violence incidents from clinical text that is already registered could be considered an important contribution to the field of personalized medicine (Ozomaro et al., 2013). In this work we therefore apply several machine learning techniques to this problem, in order to determine if prediction of violence incidents from EHR data is possible, and if so what Deep Learning or classical techniques should be applied.

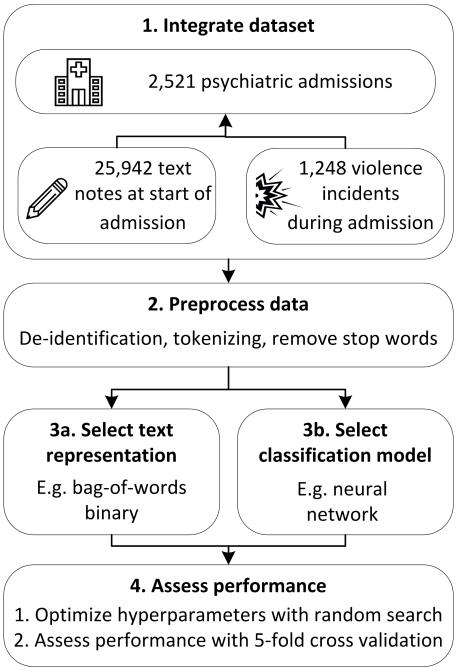


Figure 5.1: An overview of the proposed steps involved in comparing Deep Learning and classical machine learning techniques. The details of the admissions in step 1 can be seen in Table 5.1. All text representations in step 3a are visible in Table 5.2, and the classification models of step 3b are shown in Table 5.3.

5.2 Materials and Methods

In this section, we describe the used dataset, operationalize the prediction objective, elaborate on the various Deep Learning and classical techniques that are applied to the prediction problem, and describe the experimental setup. An overview of the proposed dataset and method is visible in Figure 5.1.

5.2.1 Prediction Objective

A relevant dataset for the prediction task was obtained from the Department of Psychiatry of the UMCU. This department consist of six inpatient units,

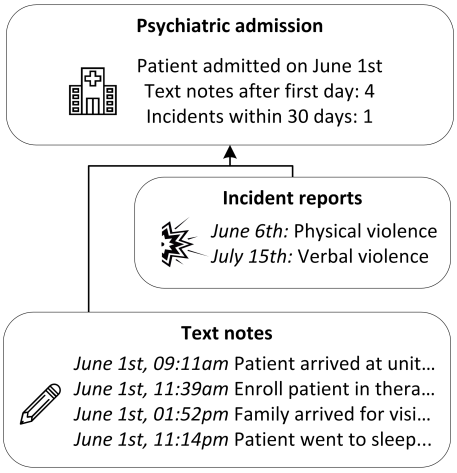


Figure 5.2: A fictional example admission is shown.

where patients with different medical histories are admitted, each with their own focus on different patient populations, diagnoses, and treatments. The department delivers secondary care to patients with severe but general psychiatric symptoms, and tertiary care to patients with more complex symptomatology or comorbidities, ensuring a diverse population. Admissions from all six units between 2013 and 2016 were included in the dataset, resulting in a total of 2,521 admissions from 1,796 unique patients, including readmissions and transfers between different units. In all six units, mandatory reporting of violence incidents by one of the health care professionals involved in the incident took place. Typically, these incidents concerned violence from patients directed at staff or at other patients, including both verbal and physical aggression. In the relevant time period, a total of 1,267 violent incidents were reported. After excluding incidents that did not involve a patient that was admitted at the time of the incident ($n = 19$), for example incidents that involved visitors rather than the patient, or incidents that happened after dismissal of a patient, a total of 1,248 incidents remained. Some descriptive statistics of the dataset per unit can be found in Table 5.1, and an example of a fictional admission is visible in Figure 5.2.

We defined the prediction objective as follows: predict for which admissions a violence incident will occur in the first 30 days, based on clinical texts that were written up to and including the first day of admission. Since in many

Table 5.1: Some descriptive statistics of the six inpatient units. An admission is classified as violent if at least one incident occurs between the second and 30th day of admission.

Unit	Population	Type of Unit	Type of Admission	No. Admissions	Violent Admissions (%)
1	Adult	Closed	Planned	307	3.6
2	Adult	Closed	Acute	1,047	7.5
3	Child, adolescent	Closed	Acute	415	13.7
4	Adolescent, adult	Closed	Planned	428	14.3
5	Child	Closed	Planned	139	34.5
6	Child	Day treatment	Planned	185	17.3

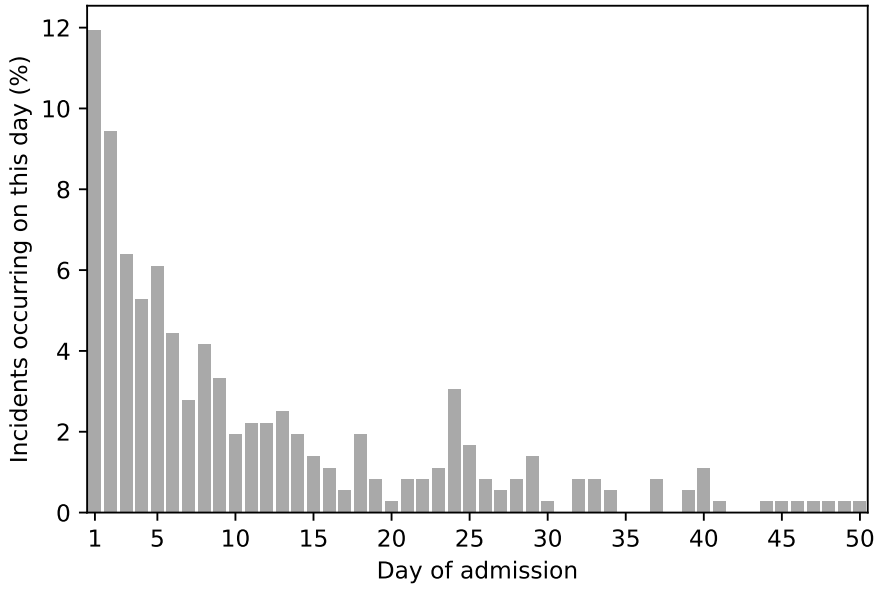


Figure 5.3: Day on which the first violence incident occurred as percentage of the total number of first incidents, cutoff after day 50. For instance, it can be verified that if violence occurs during an admission, the first incident happened on the fifth day of admission roughly 6% of the time.

admissions, relevant information was not discovered and written down until the day of admission, especially in acute admissions, the prediction task did not include violence incidents on the day of admission. Although an number of incidents that was not negligible occurred on the day of admission, exploratory analysis showed that at that point in time, insufficient data was available to make a prediction. The interval of 30 days was furthermore chosen so that the prediction was specific, while the majority of incidents were included in the prediction, given that the mean duration of admission was 40.3 days. In total, 81.9% of incidents happened during the first 30 days of admission, and it can be verified in Figure 5.3 that the amount of incidents diminished over time. To be in line with literature on this topic, we will use the Area under Curve (AUC) of the Receiver Operator Curve to report the performance of the prediction.

5.2.2 Text Dataset

Much of the clinically relevant information was entered into the EHR in free text format, either by psychiatrists or nurses. These text entries typically contain between 100 and 500 words, and are respectively referred to as doctor notes and nurse notes. The doctor note dataset comprised 11,067 notes at the start of admission, that mainly contain information such as patient history, current treatment (e.g. types of medication and therapy), and changes therein. The nurse note dataset contained 14,875 notes at the start of admission that are written three times a day by trained nurses about all admitted patients, and typically reflect the current well-being and activities of a patient. The complete corpus of doctor and nurse notes (i.e. all notes written before, during or after admission) in the same time period was also made available, totaling 1,015,931 doctor and nurse notes combined. All notes were de-identified using the De-identification Method for Dutch Medical Text (DEDUCE) (Menger et al., 2018b) before any other processing took place. The subset of notes that was available at the start of admission served as input for the prediction problem, while the entire corpus of notes were used to learn representations of text.

5.2.3 Text Representations

In order to apply classification models to these texts, a suitable numerical representation is needed. Below, we describe the representation techniques that we applied to the texts. Parameters and settings of these methods were selected based on available literature. In all representations, stop words as a single token were omitted, and where relevant the Natural Language Toolkit (NLTK) Dutch sentence tokenizer (Bird et al., 2009) was used to tokenize text. An overview of these text representations with the instantiations of their parameters is presented in Table 5.2.

First of all, the most common classical technique for representing text is a bag-of-words approach, where documents are represented as vectors of size equal to the dataset vocabulary, encoding the presence or absence of vocabulary terms. This technique that is used in many NLP and Information Retrieval (IR) applications represents documents or sentences as a multiset of their words, disregarding any grammar and word order. A possible addition to this is the use of n-grams, i.e. not only to incorporate single terms in the vocabulary, but also frequent sequences of n terms. Finally, multiplying the Term Frequency with the Inverse Document Frequency of a term, a technique usually referred to as tf-idf weighting, can be used to assess the relative im-

Table 5.2: The different types of text representations and the values of the parameters used.

Representation	Parameter	Value
Bag of words	Weighting	binary, tf-idf
	N-gram range	1–3
	No. features	1,000
Text embeddings	Level	word, document
	Model size	320
	Min frequency	50
	Epochs	20

portance of terms in a document. The performance of binary bag-of-words or bag-of-words with tf-idf weighting seems to be dependent on the type of data and type of modeling technique used (Lan et al., 2005). Since tf-idf weighting in some cases has a benefit over binary weighting, we evaluated both representations. In both cases we added bi- and trigrams to the representation, which has been demonstrated to have a positive effect on performance, while adding higher order grams can deteriorate performance (Fürnkranz, 1998). Most algorithms perform best with a number of features that is smaller than the number of examples, and using a vocabulary that is too large will lead to worse performance due to overfitting (Dalal and Zaveri, 2011; Li et al., 2004). To balance between a representation that is too small or too large, which both has negative consequences for performance, we limited the vocabulary size to the 1,000 most frequent terms.

One disadvantage of the bag-of-words approach is that information in documents is lost, such as the order of words, and negations in a sentence. Moreover, the bag-of-words vectors represent these documents by counting the frequency of words, disregarding any meaning of words or similarity between them. Recent advances in Deep Learning have been able to mitigate this problem, most notably by the introduction of the word2vec algorithm by Mikolov et al. (2013), allowing representation of text as a dense vector in a high dimensional space. The word2vec algorithm that learns embeddings for words was later extended with the paragraph2vec algorithm (Le and Mikolov, 2014) that allows representation of arbitrary-length sequences of words. We used the word2vec and the paragraph2vec algorithms to learn text embeddings, respectively on the word and document level. Before training these embedding models, text was preprocessed by mapping special characters to their ascii counterparts, transforming the text to lowercase and by removing

any remaining non-alphanumeric characters. The models were trained on the entire dataset of doctor and nurse notes, which comprised a total of just over one million texts. Only words with a minimum frequency of 50 were included, in order to filter out very uncommon words and infrequent misspellings, as well as to speed up the learning process, and to prevent overfitting. We used a typical model size of 320 nodes, and set the number of training epochs to 20. Model performance was shown rarely to decrease with increasing values for these parameters, yet the expected gain for increasing them more appeared little (Chiu et al., 2016).

5.2.4 Classification Models

As described in Section 5.2.2, the input data of the machine learning problem is a sequence of notes that is present in a patient's EHR. We applied several models to the prediction task, which requires a representation as described in Section 5.2.3 as input, either as a sequence or as a single representation. For each model we used a setup or architecture that is relatively straightforward, and is used in other literature. An overview of all models is presented in Table 5.3, along with the hyperparameters that were optimized. All other hyperparameters were fixed, for which a rationale is provided below.

The most commonly used statistical models that are applied to text classification include Neural Networks, Bayesian Classifiers, Support Vector Machines, and Decision Trees (Aggarwal and Zhai, 2012; Korde, 2012). Although pattern-based classifiers and k-nearest-neighbors type classifiers have also been applied to text classification, we did not apply them in this work because of their difficulties with imbalanced datasets, and difficulties with estimating probabilities that are needed for computing the AUC, respectively. For Neural Networks, we considered a three-layer feed-forward Neural Network, as used for example by Rajan et al. (2009). The Naive Bayes algorithm is very commonly used, and has been shown to obtain good results, for example by Deshpande et al. (2007), providing a good instantiation of a Bayesian Classifier. For Support Vector Machines, we considered a standard model with either a linear or radial kernel, which typically obtained the best results for text classification (Alsaleem, 2011; Sun et al., 2009). We used the CART algorithm finally to obtain a Decision Tree of which the depth and the number of features to consider when splitting can be controlled (Uğuz, 2011).

The models mentioned above have already proven their merits in text classification, novel Deep Learning techniques however, have recently acquired the attention of the NLP community. An additional benefit of these techniques is that they can take sequences of text as input, making use of a richer repre-

Table 5.3: An overview of Deep Learning models that take a sequence as input, and classical models that require a single input, along with the relevant hyperparameters that are tuned. Abbreviations: RNN = Recurrent Neural Network, CNN = Convolutional Neural Network, NN = Neural Network, NB = Naive Bayes, SVM = Support Vector Machine, DT = Decision Tree, GRU = Gated Recurrent Unit, LSTM = Long Short-Term Memory.

Model	**Hyperparameter**	**Range**
RNN	Learning rate	10^{-5}–10^{-1}
	Cell type	GRU, LSTM
	Layer size	2^4–2^8
	Dropout rate	0.5–0.9
CNN	Learning rate	10^{-5}–10^{-1}
	No. filters	2^4–2^8
	Filter size	3–7
	Pooling size	2–7
	Dropout rate	0.5–0.9
	Fully connected layer size	2^4–2^8
NN	Learning rate	10^{-5}–10^{-1}
	Layer size	2^4–2^8
	Regularization constant	10^{-5}–10^{-1}
NB	N/a	–
SVM	Gamma	10^{-5}–10^{-1}
	C	10^{-5}–10^{-1}
	Kernel	linear, radial
DT	Max depth	2^1–2^4
	Max features	0.25–0.75
	Min samples split	2^1–2^4

sentation of the input. Firstly, Recurrent Neural Networks (RNNs) work by processing a sequence of inputs one-by-one, adjusting its internal state at each step. Based on each input, an output is computed, which serves as input for the next step along with the next item in the input sequence. In our case, only a classification label is desired as output, making a many-to-one RNN setup appropriate. In this setup the sequence of texts is processed sequentially, providing an internal encoding of the input, after which the binary outcome is determined based on this encoding. In our setup a unidirectional RNN with dropout regularization was used, and we instantiate the recurrent cell of the RNN with both a Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) cell, applied for example by Liu et al. (2017) and Tang et al. (2015). Secondly, Convolutional Neural Networks (CNNs) compute a similar mapping in a different way. CNNs are most commonly applied to image or video data (i.e. sequences of two or three dimensions), yet they are also applicable to text data (a sequence of one dimension). They work by applying a sliding window to a sequence using the convolutional operator, which is able to automatically learn higher-order features. This is done by sliding a kernel with randomly initiated weights over the sequence, after which it automatically updates its weights to learn from each example. One possible advantage of CNNs over RNNs is that in RNNs, the last items in a sequence have a relatively high impact on the outcome, while CNNs do not exhibit such a bias. We used a straightforward setup with a Convolutional layer, followed by a Dropout layer and a Max Pooling layer, after which a single fully connected layer determined the output from the previous layers, such as used in Hughes et al. (2017) and Zhang et al. (2015). Both the RNN and the CNN were trained for 50 epochs.

5.2.5 Experimental Setup

All text representations in Table 5.2 were combined with all classification models in Table 5.3 to experimentally evaluate the performance of each pair. To be able to apply classical models to sequential data, the input sequences needed to be aggregated into a single vector. For the bag-of-words approaches, all texts in the sequence were concatenated, after which the bag-of-word features were computed. For the word and document embeddings, the sequence of vectors was averaged to compute a single vector for each instance. We used 5-fold stratified cross validation to compute the performance of each combination, measured by the AUC, along with its standard deviation. For finding the optimal hyperparameters we used a random search, as suggested by Bengio (2012). The hyperparameters were sampled from either a uniform or a log uniform distribution with a typical range that can be seen in Table 5.3, and a

total of 250 random parameter samplings were used for each evaluation.

5.3 Results

The results of the experimental evaluation can be seen in Table 5.4, and the optimal hyperparameters have been included in Supplementary Materials Table 5.5 for completeness. It can be seen that the best prediction was obtained by combining Document Embeddings with a Recurrent Neural Network, closely followed by Binary Bag of Words with a Recurrent Neural Network and Document Embeddings with a Support Vector Machine.

From the perspective of the different representations, the weighting scheme of the Bag-of-words approach did not result in a clear difference in performance between different models, and neither was there a strong difference in applying the text embeddings on the word level or the document level seen. Overall however, text embeddings resulted in better performance than the bag-of-words approaches.

Of the classical models, the Naive Bayes algorithm, which despite its simplicity often yields good results in text classification, is in this case not among the top performing algorithms. Decision Trees were able to perform slightly better, but also could not match the performance of the other algorithms. A possible explanation for this is that these models were relatively simple, and they were not able to find the relatively complex patterns that were needed to accurately assess violence risk. The Neural Network algorithm showed better results, especially in combination with the text embeddings, and a clear difference between Bag-of-words representations and Text embedding representations could be seen. The Support Vector Machine finally was able to predict violence incidents in the first 30 of admission days best of the four classical models, with just a marginal difference in performance over different representations of text.

For the two Deep Learning models, it can be seen that especially the RNN outperformed several of the classical models. RNNs outperformed CNNs for this classification task as well, with a similar margin over different text representations. The AUC scores were however among the lowest when Word Embeddings were used for representing text, while these did not cause a decrease in performance for the classical models, suggesting that they did contain the information that was needed to assess the violence risk. Despite the fact that a sequence of Word Embeddings is a richer representation than the average over these Word Embeddings, the fact that the length of the input sequence was no longer proportional to the number of instances might provide difficul-

Table 5.4: The performance for optimal hyperparameter values for each of the representations combined with the models, based on a 5-fold stratified cross validation. The performance is measured in AUC, along with its standard deviation. The best performance over different models is marked with an ^{a}, the best performance over representations with a ^{b}. Abbreviations: RNN = Recurrent Neural Network, CNN = Convolutional Neural Network, NN = Neural Network, NB = Naive Bayes, SVM = Support Vector Machine, DT = Decision Tree.

Model	**Bag-of-Words Binary**	**Bag-of-Words tf-idf**	**Word Embeddings**	**Document Embeddings**
RNN	0.771 ± 0.018^{b}	0.753 ± 0.031	0.654 ± 0.043	$0.788 \pm 0.018^{a,b}$
CNN	0.729 ± 0.030	0.716 ± 0.038	0.684 ± 0.038	0.763 ± 0.024^{a}
NN	0.727 ± 0.033	0.717 ± 0.038	0.751 ± 0.036^{a}	0.745 ± 0.022
NB	0.686 ± 0.026	0.704 ± 0.034^{a}	0.700 ± 0.051	0.692 ± 0.046
SVM	0.759 ± 0.040	0.756 ± 0.036^{b}	0.764 ± 0.024^{b}	0.770 ± 0.029^{a}
DT	0.727 ± 0.018^{a}	0.719 ± 0.041	0.685 ± 0.041	0.665 ± 0.035

ties in practice. In this case, given the computational resources and dataset size, the classical models were able to handle the condensed word embedding averages better. Both Deep Learning models performed best when combined with Document Embeddings, resulting in the overall optimal performance of 0.788 AUC, followed by the two bag-of-words representations.

5.4 Discussion

The results of the experimental evaluation in Table 5.4 show the best result was obtained by combining Document Embeddings with a Recurrent Neural Network, although the difference with different methods was relatively small, and in some cases, smaller than the standard deviation. Despite these small differences, it can also be seen that most of the top performing methods either used text embeddings for representing text, or a Deep Learning model for classification. Although Deep Learning techniques did not exclusively lead to good classification results, they did give a small but consistent advantage in performance. In most research, Deep Learning methods are superior in large datasets, this research shows that Deep Learning methods can even start to outperform classical methods in modestly sized datasets. While near-optimal performance can be achieved with a bag-of-words approach combined with Support Vector Machines, applying Deep Learning techniques to clinical text datasets of this size will be especially beneficial in performance-critical applications. One advantage of using classical techniques on the other hand is the reduction in training time compared to Deep Learning techniques. Although the training time of Deep Learning techniques on a dataset of this size is not inhibitive, training a SVM, for instance, can in this case be up to an order of magnitude faster than training a RNN. The difference in classification time was negligible. For the Deep Learning models on the other hand, no optimization of the network architecture was done, and a relatively large number of hyperparameters was optimized with a constant number of hyperparameter samplings. Additional computational resources and experimentation with network setup might further improve the performance of the Deep Learning models, while a similar gain is not expected for classical techniques.

One possible limitation of the experimental evaluation was the 5-fold cross validation strategy for validating the model performances. Since several combinations of text representations, classification models, and parameter settings have been evaluated, some degree of overfitting cannot be prevented. As a result, a small bias may exist in the optimal outcome of 0.788 AUC, which does not influence the comparison, but does inhibit regarding this as a definitive

result for the health care practice. Additional research is needed to precisely establish to what level of accuracy risk assessment can be performed in an automated way using clinical text. On the other hand, adding structured variables from the EHR, such as medication use, diagnosis, and patient demographics, as well as increasing the sample size are future research directions that may be able to further improve the result.

The results of the performed experiments finally have some implications for the assessment of violence risk in the psychiatric practice as well. As described in the introduction, the most commonly used violence risk assessment instruments show a median AUC between 0.70 and 0.74 when measured in a meta study over different sites (Singh et al., 2011). Although higher AUC scores have been reported in individual studies, lower scores have been reported as well, indicating that the performance of these assessment instruments is not very generalizable to other patient populations or health care institutions. Our machine learning approach, which achieved an optimal AUC of 0.788 on a patient sample size that is comparable to the combined meta study sample sizes for each of these instruments, shows that assessing violence risk from clinical text in the EHR is a very promising approach. The main advantages of this approach over existing risk assessment tools is that the assessment can be specifically tailored to the population of an institution, and that it can constantly be adjusted over time. This also allows measuring the performance of the method on the relevant clinical population. The assessment can furthermore be automatically performed based on already available clinical data, thus saving time and cost without sacrificing assessment accuracy or imposing significant changes in the clinical process. Although the experiment setup necessitates some further research to fully establish to which extent prediction of violence using EHR data is possible, our research shows that this approach is promising, and that in the future it can provide an important novel addition to the field of violence risk assessment.

5.5 Conclusions

Violence during psychiatric admissions is a problem that causes a high burden for both patients and hospital staff. Although several of its associated individual factors are known, and structured risk instruments for assessing the risk are available, meta studies reveal that generalizing these individual factors or instruments to other populations is not always straightforward. In this work, we investigated whether automatic assessment of violence risk is possible with textual data that is already captured in an EHR. To do so, we

compared classical machine learning techniques and Deep Learning techniques. For the study we used a novel and previously unexplored dataset of the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in The Netherlands.

Our experiment evaluated all combinations of a text representation (Bag-of-words with binary weighting, Bag-of-words with tf-idf weighting, Word embeddings, or Document embeddings) and a classification model (Recurrent Neural Network, Convolutional Neural Network, Neural Network, Support Vector Machine, Naive Bayes, or Decision Tree). We used random search with 5-fold cross validation for optimizing hyperparameters. The results of our evaluation show that the best result is obtained by combining Document Embeddings with a Recurrent Neural Network ($\text{AUC} = 0.788 \pm 0.018$), closely followed by a Binary Bag of Words combined with a Recurrent Neural Network ($\text{AUC} = 0.771 \pm 0.018$), and Document Embeddings combined with a Support Vector Machine ($\text{AUC} = 0.770 \pm 0.029$).

A relatively small but consistent improvement in performance could be seen for Deep Learning techniques over classical machine learning techniques. Deep Learning techniques furthermore have the advantage of allowing more additional experimentation with the model setup, while on the other hand, the training time of classical machine learning techniques can be up to one order of magnitude smaller. Using Deep Learning techniques on a dataset of this size and for this type of problem therefore shows promise, especially in performance critical applications.

Our results finally have potential implications for the psychiatric practice as well, although the exact accuracy of automatic risk assessment from EHR data needs to be established in further research. The results we obtained are improvements over the median AUC of structured risk assessment instruments as measured in meta studies over different sites, with a comparable sample size. The proposed method using EHR data is furthermore customizable to a specific population or institution, circumventing the problem of generalization. Automatic assessment of violence risk therefore is a promising approach that can in the future provide an important addition to the psychiatric practice.

5.6 Supplementary Materials

Table 5.5: An overview of the optimal hyperparameters for each combination of a text representation and a classification model. RNN = Recurrent Neural Network, CNN = Convolutional Neural Network, NN = Neural Network, NB = Naive Bayes, SVM = Support Vector Machine, DT = Decision Tree, GRU = Gated Recurrent Unit, LSTM = Long Short-Term Memory.

Model	**Hyperparameter**	**Bag-of-Words Binary**	**Bag-of-Words tf-idf**	**Word Embeddings**	**Document embeddings**
RNN	Learning rate	$4.1 \cdot 10^{-2}$	$5.3 \cdot 10^{-2}$	$8.3 \cdot 10^{-3}$	$4.6 \cdot 10^{-2}$
	Cell type	LSTM	LSTM	GRU	LSTM
	Layer size	193	63	185	129
	Dropout rate	0.8	0.7	0.5	0.8
CNN	Learning rate	$1.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$4.24 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$
	No. filters	41	69	45	49
	Filter size	3	3	3	4
	Pooling size	5	6	5	6
	Dropout rate	0.9	0.7	0.5	0.9
	Fully connected layer size	18	36	121	85
NN	Learning rate	$1.5 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$6.4 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$
	Layer size	172	30	36	254
	Regularization constant	$4.7 \cdot 10^{-4}$	$2.2 \cdot 10^{-4}$	$7.1 \cdot 10^{-2}$	$9.9 \cdot 10^{-2}$
NB	N/a	–	–	–	–
SVM	C	0.40	0.50	2.52	0.40
	Gamma	$3.1 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$3.6 \cdot 10^{-4}$	$7.9 \cdot 10^{-4}$
	Kernel	radial	radial	radial	radial
DT	Max depth	2	3	2	4
	Max features	0.52	0.47	0.56	0.84
	Min samples split	5	3	11	3

6 | Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records

Importance: Inpatient violence remains a significant problem in psychiatry despite existing risk assessment methods. Current assessments' lack of robustness and high effort of use in practice might be mitigated by utilizing routinely registered clinical notes.

Objective: To develop and validate a multivariable prediction model for assessing inpatient violence risk, based on machine learning techniques applied to clinical notes written in patients' Electronic Health Records.

Design, Setting and Participants: In a prognostic study, we used retrospective clinical notes registered in EHRs during admission, in two independent psychiatric health care institutions. No exclusion criteria for individual patients were defined. In site 1, all adult admissions between January 2013 and August 2018 were included, and in site 2 all admissions to general psychiatric wards between June 2016 and August 2018 were included. Data were analyzed between September 2018 and February 2019.

Main Outcomes and Measures: Predictive validity and generalizability of prognostic models, measured using Area Under the Curve (AUC).

Results: Clinical notes recorded during a total of 3,189 admissions of 2,209 individuals in site 1 (mean [SD] age 34.0 [16.6]; 48.2% male) and 3,253 admissions of 1,919 individuals in site 2 (mean [SD] age 45.9 [16.6]; 65.8% male) were analyzed. Violent outcome was determined using the Staff

Observation Aggression Scale Revised (SOAS-R). We used nested cross validation to train and evaluate models that assess violence risk during the first four weeks of admission, based on clinical notes available after 24 hours. The predictive validity of models was measured both in site 1 (AUC = 0.80; 95% CI 0.77 to 0.82) and site 2 (AUC = 0.76; 95% CI 0.73 to 0.80). Applying trained models to other-site data resulted in a significantly lower AUC at the $\alpha = 0.01$ level in site 1 (AUC difference = 0.08; 95% CI 0.05 to 0.11; $p < 0.001$) and site 2 (AUC difference = 0.12; 95% CI 0.09 to 0.16; $p < 0.001$).

Conclusions and relevance: Internally validated predictions result in AUCs with good predictive validity, showing that automatic violence risk assessment using routinely registered clinical notes is possible. Other-site validation of trained models corroborates previous findings that violence risk assessment generalizes modestly to different populations.

This work was originally published as:

Menger, V., Spruit, M., van Est, R., Nap, E., and Scheepers, F. (2019c). Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Network Open*, 2(7):e196709

6.1 Introduction

Violence in psychiatric inpatient wards remains a significant problem. A study combining data from 35 sites worldwide shows 14–20% of patients commit at least one act of violence during inpatient treatment (Iozzino et al., 2015), and surveys consistently show the vast majority of practitioners being affected by violence at some point during their career (van Leeuwen and Harte, 2017). Adverse effects on both patients' and caregivers' well-being, such as injury, low morale, and high absentee levels, are well known (Inoue et al., 2006; Nijman et al., 2005).

As an important part of managing inpatient violence, structured Violence Risk Assessment (VRA) instruments have been proposed based on a combination of static and dynamic risk factors. Their predictive validity surpasses that of unstructured clinical judgment, and a reasonable adoption in practice has been achieved, with over half of all risk assessments performed using an instrument (Singh et al., 2014). Meta analyses however reveal that only a small subset of risk factors for violent behavior generalize to different populations (Dack et al., 2013; Steinert, 2002), and VRA instruments are consequently limited by the robustness of the individual factors that comprise them (Singh et al., 2011; Yang et al., 2010). In addition, the time needed to perform a structured assessment, ranging from minutes to hours, has been identified as an obstacle for daily practice. Although adopting a VRA instrument diminished the number of violence incidents in one RCT (Abderhalden et al., 2008), other research suggests its benefits in practice are still moderate due to its limitations (Viljoen et al., 2018; Wand, 2011).

Developing a prognostic model based on textual data registered in patients' EHRs might offer a novel approach to improve violence risk assessment. The fact that these data are unstructured, and originally designated for treatment, presents methodological challenges, but also opportunities in combating selection bias and exploring new associations (Menger et al., 2016). Machine learning, a term that refers to a set of statistical techniques that learn from large and potentially noisy datasets, is eminently well suited for this kind of task. Prognostic models obtained using these techniques are automatically tailored to the relevant population, and can be fitted in the care process without imposing additional administrative load, circumventing drawbacks of structured VRA instruments. Although many fields of medicine have seen convincing cases of algorithms aiding clinical decision making (e.g. cardiology (Weng et al., 2017), dermatology (Esteva et al., 2017), oncology (Kourou et al., 2015)), the field of psychiatry still seems only on the verge of transforming in

this direction (Fernandes et al., 2017; McIntosh et al., 2016).

In this prognostic study, we tested for the first time to what extent textual data from the EHR can be used to automatically assess violence risk, by developing and validating multivariable prediction models, based on routinely collected clinical notes from two independent psychiatric treatment centers.

6.2 Methods

In this study, we used data that were extracted from EHRs of two independent psychiatric treatment centers. Data sources were not connected to each other or to sources outside the separate hospitals. We used de-identified datasets, by de-identifying clinical notes using the DEDUCE method (Menger et al., 2018b). Demographical variables were limited to gender, year of birth, and DSM diagnosis. The study was reviewed, and approved by the University Medical Center Utrecht ethical committee. The committee assessed that obtaining informed consent retroactively was not necessary, because of the retrospective nature of the study, the amount of participants, the fact that no extra data were obtained, and the use of de-identified data. For reporting this study, we followed the guidelines of Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and Reporting Guidance for Violence Risk Assessment Predictive Validity Studies (RAGEE) (Collins et al., 2015; Singh et al., 2015).

6.2.1 Cohort Definition

Site 1, used for internal method validation, is the psychiatry department of the academic medical center in Utrecht, the Netherlands. It delivers both secondary and tertiary care in four closed short-term treatment wards, among which an acute ward, and wards that focus on treatment of patients with psychotic disorders, mood disorders, and developmental disorders. A new admission was registered both when a new patient was admitted, and when a patient was transferred between psychiatric wards. We tolerated an absence of at most two weeks during admission—for example due to discharge and readmission, or temporary admission in a non-psychiatric department—longer absences were registered as a new admission as well. Admissions in the developmental disorder ward were excluded, based on age and nature of violence. All admissions in other wards that started between January 2013 and August 2018 were included in the dataset. We defined no exclusion criteria based on diagnosis, co-morbidity, or other psychopathological conditions, to maximize

translational value of predictive models. The resulting dataset consisted of 3,201 admissions of 2,211 unique patients.

Site 2, used for external method validation, is a general psychiatric hospital that delivers secondary care, with an additional focus on addiction care. It consists of 47 treatment wards in the area of Rotterdam, the Netherlands. To match the original dataset, admissions to forensic psychiatric wards (n=2), long-term care wards (n=25), and wards that exclusively offer addiction care (n=9) were not included in the study. All admissions in the 11 retained wards that started between June 2016 and August 2018 were included in the dataset. Other conditions were kept equal. The resulting dataset consisted of 3,277 admissions of 1,937 unique patients. How datasets in both sites were extracted from EHR systems, and how data quality is secured, is detailed Supplementary Materials, Section 6.5.1. We did not merge datasets, but used the dataset of site 1 for developing a machine learning approach, and then used the dataset of site 2 for externally validating this approach, and finally exchanged trained models between the sites.

6.2.2 Data Selection

Clinical notes that were written by psychiatrists and nurses were directly extracted from patients' EHRs. We hypothesize that free text contains information that cannot easily be captured in structured form, e.g. behavioral cues or social interactions, yet is relevant for violence risk assessment. Notes that were written in the four weeks before admission up to the first 24 hours of admission were included in the datasets. Admissions with less than 100 words registered after 24 hours (12 and 24 admissions in site 1 and site 2 respectively), were excluded from the dataset.

6.2.3 Outcome Variable

Reports of violence incidents were used to determine the outcome for each admission. In both sites mandatory reporting of all violence incidents takes place, including patient-staff and patient-patient violence. In the incident form, staff members that were involved in the incident were required to fill in structured information, a textual description of the incident, and incident severity as measured by the Staff Observation Aggression Scale Revised (SOAS-R) (Nijman et al., 1999). Our definition of a violence incident included all threatening and violent behavior of a verbal or physical nature directed at another person, but excluded self-harm and inappropriate behavior such as substance abuse, sexual intimidation, or vandalism. A positive outcome was defined as presence

of at least one incident in the first four weeks of admission, excluding the first 24 hours. No distinction in incident severity was made.

6.2.4 Exploratory Analysis

To examine the potential predictive power hidden in clinical notes, we extracted the 1000 most frequent terms in the clinical notes, including bigrams, as binary variables. We then assessed the strength of each term's association with the outcome using a chi-squared test, and computed Matthews correlation coefficients to obtain the direction of the association. We selected the top 10% of predictors based on their chi-squared scores in 1000 repeated samples with replacement, computing the fraction of times a term was included among the top predictors as a measure of within-dataset generalizability of predictors.

6.2.5 Machine Learning Models

We used a machine learning approach to perform violence risk assessment. Machine learning algorithms are able to detect patterns, if present, in historical data, based on which a prediction of the future course of treatment can be made. Such an approach applied to textual data must comprise two steps: transforming clinical notes into a suitable numerical representation, and subsequently feeding these numerical representations into a classification model.

To transform the clinical notes into a numerical form, we used the novel paragraph2vec algorithm that learns an accurate numerical representation from a large corpus of text in an unsupervised way, i.e. unrelated to outcome (Le and Mikolov, 2014). This algorithm, founded in deep learning theory, is capable of using not only verbatim words in a text to determine a representation, but also word order and the context of words such as negations. In previous work, we have shown the potential added value of this technique over a traditional bag-of-words approach when applied to violence risk assessment (Menger et al., 2018a). The model was trained using a large internal set of clinical notes (i.e. not only notes relevant for assessment), with model settings based on available literature without optimization (see Supplementary Materials, Section 6.5.2 and Table 6.5 for details)

The numerical representations of texts were subsequently fed into a Support Vector Machine with a radial kernel (Cortes and Vapnik, 1995), a model that has previously been shown as appropriate for text classification (Joachims, 1998). It works by first mapping data points to a higher-dimensional space, and by then inferring a decision boundary that maintains a maximum margin

to these data points. New data points are subsequently classified according to which side of the boundary they lie on.

6.2.6 Statistical Analysis

Model training and estimation of model predictive validity were done in a nested cross validation setup, ensuring that admissions used for learning models were never used to simultaneously determine predictive validity. Different admissions of the same patient were additionally never split over different folds, to ensure that predictions were not influenced by information from future admissions from the same patient. The final AUC was computed by averaging the AUCs of the five outer cross validation folds, while confidence intervals and standard error of the mean were established using the method of DeLong et al. (1988) (Hajian-Tilaki and Hanley, 2002). Additionally, performance metrics such as sensitivity, specificity, and relative risk were computed by pooling predictions over folds (Forman and Scholz, 2010). Experimental setup is detailed further in Supplementary Materials, Section 6.5.3. After finalizing the results in site 1, an external validation of the machine learning approach was performed in site 2, by training a new model with equal experimental setup. To further elucidate model performance, we investigated predictive validity for early violence vs late violence and short vs long admission subgroups. Finally, trained models were exchanged between sites to test their generalizability.

For the tokens discovered in exploratory analysis, the association with the outcome was determined using a chi-squared test with a Holm-Bonferroni correction to control the family wise error rate. Differences in AUC between various internal and external validations were tested for significance using the method of DeLong et al. (1988) (Robin et al., 2011). We used a paired test when comparing two models on the same dataset, i.e. when comparing the cross validated assessment and assessment using a pre-trained model, to account for correlation between the two AUCs. In all other cases we used an unpaired test. All statistical significances in this study were assessed using two-sided tests at the $\alpha = 0.01$ level. The code for machine learning and statistical analysis was developed in Python (version 3.6), and is made publicly available (see Supplementary Materials, Section 6.5.4).

6.2.7 Qualitative Evaluation

After finalizing method and results in both sites, a qualitative evaluation was conducted in a focus group with participants including practitioners, data analysts, and patient representatives from both sites. Participants discussed

the method as presented by one of the researchers, and interpreted the results. The participants' attitude towards the method was positive, and its translation between sites was deemed appropriate. No changes were introduced to the study as a result of the focus group.

6.3 Results

6.3.1 Datasets

The final datasets (Table 6.1) consist of 3,189 admissions from 2,209 patients in site 1, and 3,253 admissions from 1,919 patients in site 2. Populations differ in age (mean age 34.0 years and 45.9 years), gender (48.2% and 65.8% male), and distribution of diagnoses. In both sites the most commonly occurring diagnosis was schizophrenia or other psychotic disorders, followed by a mood disorders and personality disorders in site 1, and followed by substance-related disorders and bipolar disorders in site 2. Similar lengths of stay (16.0 and 15.0 days), median length of clinical notes (2,091.0 and 1,961.0 words), and admissions with a positive outcome (9.1% and 7.7%) were registered in both sites.

Table 6.1: Descriptive statistics of the datasets obtained in the two sites. Demographics and DSM diagnosis percentages are relative to the number of admissions. Abbreviations: SD = Standard Deviation, IQR = Inter Quartile Range, SOAS-R = Staff Observation Aggression Scale Revised, DSM = Diagnostic and Statistical Manual of Mental Disorders.

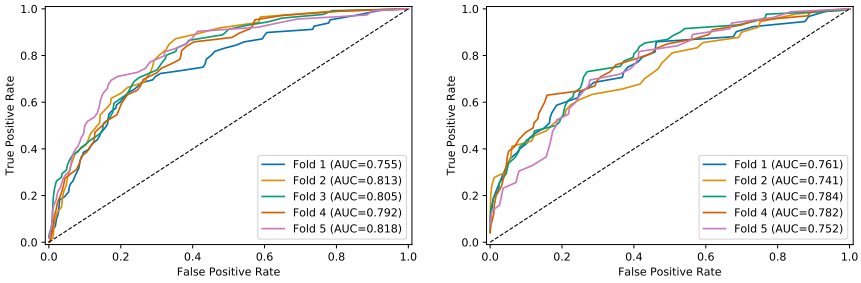
	Site 1	**Site 2**
Demographic		
Age, mean (SD)	34.0 (16.6)	45.9 (16.6)
Men, No. (%)	1,536 (48.2%)	2,097 (65.8%)
Dataset		
Admissions, No.	3,189	3,253
Unique patients, No.	2,209	1,919
Length of stay, median days (IQR)	16.0 (6.0–41.0)	15.0 (5.0–40.5)
Number of words in notes, median (IQR)	2,091.0 (1,541.0–2,981.0)	1,961.0 (1,160.0–3,060.0)
Admissions with positive outcome, No. (%)	290 (9.1%)	247 (7.7%)

Table 6.1 (continued): Descriptive statistics of the datasets obtained in the two sites.

	Site 1	Site 2
Incidents		
Incidents during admission, No.	962	652
Incidents during first 4 weeks, No. (%)	658 (68.4%)	318 (48.8%)
Incidents during first 24 hours, No. (%)	90 (9.4%)	42 (6.4%)
SOAS-R score, median (IQR)	12.0 (8.0–16.0)	11.0 (7.0–14.0)
DSM diagnosis, No. (%)		
Anxiety disorder	92 (2.9%)	63 (2.0%)
Bipolar disorder	65 (2.0%)	170 (5.3%)
Delirium, dementia, amnesia, and other cognitive disorders	20 (0.6%)	109 (3.4%)
Depressive disorder	106 (3.3%)	150 (4.7%)
Developmental disorder	180 (5.6%)	29 (0.9%)
Eating disorder	57 (1.8%)	10 (0.3%)
Mood disorder	580 (18.2%)	10 (0.3%)
Personality disorder	214 (6.7%)	116 (3.6%)
Substance-related disorder	99 (3.1%)	373 (11.7%)
Schizophrenia or other psychotic disorder	860 (27.0%)	685 (21.5%)
None within 12 weeks	795 (24.9%)	1,392 (43.7%)
Other	121 (3.8%)	146 (4.6%)

6.3.2 Machine Learning Models

Several performance metrics of predictive validity, both for in-site validation using nested cross-validation, and for other-site validation of pre-trained models were computed (Table 6.2). Optimal hyperparameters are shown in Supplementary Materials, Table 6.6. An optimal AUC of 0.797 (95% CI 0.771 to 0.822) is achieved for the internal validation of the method in site 1, while the optimal AUC for the external validation of the method in site 2 is 0.764 (95% CI 0.732 to 0.797) (Figure 6.1). The difference in internal cross validation AUC between the two sites is not significant at the $\alpha = 0.01$ level



(a) Site 1 (b) Site 2

Figure 6.1: Receiver Operator Characteristics (ROCs) for internal cross validations. ROCs are shown for each fold, based on internal cross-validation in site 1 (a) and site 2 (b). AUC = Area Under Curve.

(AUC difference = 0.032; 95% CI -0.009 to 0.074; $p = 0.12$). Specificity (i.e. prediction in the negative class) of models is higher (0.935–0.947) than sensitivity (i.e. prediction in the positive class; 0.334–0.336). The relative risk of violent outcome for patients with predicted high risk vs low risk is 5.121 and 6.297 in site 1 and site 2 respectively.

The validation of pre-trained models in the other site resulted in AUCs of 0.722 (95% CI 0.690 to 0.753) in site 1, and 0.643 (95% CI 0.610 to 0.675) in site 2. The difference in AUC between the internally trained model, and model trained on other-site data is significant both in site 1 (AUC difference = 0.075; 95% CI 0.045 to 0.105; $p < 0.001$) and site 2 (AUC difference = 0.121; 95% CI 0.085 to 0.156; $p < 0.001$) at the $\alpha = 0.01$ level. While specificity is still similar, both sensitivity and relative risk are lower as well compared to in-site validations.

We finally examined model performance in assessing early vs late violence, and violence during short vs long admissions. In both internal and external validations, and in both sites, predictive validity is higher for early violence than for late violence, and higher for short admissions than for long admissions. However, the difference was never significant at the $\alpha = 0.01$ level. For example, for the internal validation in site 1, the difference in AUC for assessing early violence vs late violence was 0.046 (95% CI -0.003 to 0.094, $p = 0.06$), and the difference in AUC for assessing violence during short admissions vs long admissions was 0.012 (95% CI -0.041 to 0.066, $p = 0.65$). Full subgroup analysis is included in Supplementary Materials, Tables 6.7 and 6.8.

Table 6.2: Predictive validity of prognostic models, in both sites, and both internally and externally trained. Abbreviations: CV = Cross Validation, AUC = Area Under the Curve, CI = Confidence Interval, SE = Standard Error of the mean

Evaluation	**Internal (CV), site 1**	**Internal (CV), site 2**	**External model, site 1**	**External model, site 2**
Model evaluated in site	1	2	1	2
Model trained in site	1	2	2	1
AUC [SE] (95% CI)	0.797 [0.013] (0.771–0.822)	0.764 [0.017] (0.732–0.797)	0.722 [0.016] (0.690–0.753)	0.643 [0.017] (0.610–0.675)
Admissions, No.	3,189	3,253	3,189	3,253
True Negative	2,711	2,847	2,682	2,793
False Negative	193	164	218	214
True Positive	97	83	72	33
False Positive	188	159	217	213
Specificity (95% CI)	0.935 (0.930–0.940)	0.947 (0.943–0.951)	0.925 (0.921–0.930)	0.929 (0.926–0.933)
Sensitivity (95% CI)	0.334 (0.287–0.383)	0.336 (0.285–0.389)	0.248 (0.205–0.296)	0.134 (0.097–0.179)
Relative Risk (95% CI)	5.121 (4.109–6.330)	6.297 (4.956–7.922)	3.314 (2.581–4.214)	1.885 (1.305–2.673)

6.3.3 Exploratory Analysis

Out of the 1,000 most frequent terms from clinical notes, the top twenty terms by generalizability within dataset were selected (Table 6.3 and Table 6.4). Several terms such as *aggressive*, *angry*, *verbal*, *threatening*, and *irritated* can directly be linked to violence, while other terms such as *reacts*, *walks*, and *speaks* describe behavioral cues that may indirectly be related to violence. The terms *aggressive* and *walked*, and their synonyms are seen in both sites. Other terms do not directly co-occur in both sites, but have a counterpart with a similar meaning (e.g. *colleague* vs *staff*, and *door* vs *office*). All terms generalize well within the dataset, being chosen among the top 10% in repeated sampling at least 95% of the time (Ratio). In site 1 the terms *aggressive*, *reacts*, and *offered* generalize best within the dataset, while in site 2 the terms *verbal*, *threatening*, and *aggression* comprise the top 3. In site 1 and site 2 respectively the 47 and 21 terms with highest χ^2-score are significantly associated with the outcome at the $\alpha = 0.01$ level after applying a Holm-Bonferroni correction (P_{corr}). Matthews correlation coefficients range from -0.14 to 0.17, showing weak correlations (MCC). Most terms have a positive correlation with violent outcome, except *status voluntary* and *dejected* in site 1, which are negatively correlated with violent outcome.

6.4 Discussion

As far as we know, this is the first time that readily available clinical notes from patients EHRs are used to assess inpatient violence risk. We applied machine learning techniques to retrospective textual data, in order to train a model that differentiates patients who show violent behavior during the first four weeks of admission. Until now, no study has performed VRA using clinical text, and no study has tested automatic VRA in multiple sites. Especially the AUCs of internally cross validated predictions (0.80 and 0.76) lie in the range that can be seen as acceptable for application in practice. While in-site validation of models obtained good results, other-site validation of pre-trained models resulted in significantly lower predictive validity, corroborating previous findings that VRA generalizes modestly over different populations. This strengthens the case for using locally developed and/or trained models and methods for violence risk assessment. Our choice to balance between false positives and false negatives for reporting outcomes resulted in higher predictive validity in the low-risk class (e.g. sensitivity) than in the high-risk class (e.g. specificity), which is largely in line with existing VRA research. Unfortunately, until now

Table 6.3: Results of exploratory analysis in Site 1. We extracted the top 20 terms by within-dataset generalizability (Ratio). Terms are written in Dutch, additionally literal English translations of terms using the Van Dale Dutch–English dictionary, 3rd edition, are provided between parentheses. Matthews correlation coefficient is computed to assess the direction of association between the term and outcome (MCC). Statistical significance is assessed using a χ^2 test, a Holm-Bonferroni correction is applied to obtain corrected p-values ($P_{corr}(\chi^2)$). ** indicates statistical significance at the $\alpha = 0.001$ level, * indicates significance at the $\alpha = 0.01$ level. Abbreviations: MCC = Matthews Correlation Coefficient, P_{corr} = corrected p-value.

Rank	**Term**	**Ratio**	**MCC**	$P_{corr}(\chi^2)$
1	agressief ('aggressive')	1.00	0.17	<0.001**
2	reageert ('reacts')	1.00	0.15	<0.001**
3	aangeboden ('offered')	1.00	0.14	<0.001**
4	boos ('angry')	1.00	0.16	<0.001**
5	deur ('door')	1.00	0.14	<0.001**
6	loopt ('walks')	1.00	0.15	<0.001**
7	ibs ('arrest')	1.00	0.14	<0.001**
8	aanbieden ('offer')	1.00	0.12	<0.001**
9	noodmedicatie ('emergency medication')	0.99	0.14	<0.001**
10	liep ('walked')	0.99	0.12	<0.001**
11	agressie ('aggression')	0.99	0.13	<0.001**
12	vraagt ('asks')	0.99	0.13	<0.001**
13	status vrijwillig ('status voluntary')	0.99	-0.12	<0.001**
14	psychotisch ('psychotic')	0.98	0.12	<0.001**
15	collega ('colleague')	0.98	0.11	<0.001**
16	spreekt ('speaks')	0.97	0.12	<0.001**
17	gehouden ('obliged')	0.97	0.11	<0.001**
18	beoordelen ('judge', verb)	0.96	0.11	<0.001**
19	momenten ('moments')	0.96	0.12	<0.001**
20	somber ('dejected')	0.95	-0.14	<0.001**

Table 6.4: Results of exploratory analysis in Site 2. See Table 6.3 for explanation.

Rank	**Term**	**Ratio**	**MCC**	$P_{corr}(\chi^2)$
1	verbal ('verbal')	1.00	0.14	<0.001**
2	dreigend ('threatening')	1.00	0.13	<0.001**
3	agressie ('aggression')	1.00	0.15	<0.001**
4	hierop ('(up)on this')	1.00	0.13	<0.001**
5	kantoor ('office')	1.00	0.12	<0.001**
6	personeel ('staff')	1.00	0.12	<0.001**
7	aangesproken ('spoke to')	1.00	0.11	<0.001**
8	agressief ('aggressive')	0.99	0.11	<0.001**
9	gevaar agressie ('danger aggression')	0.99	0.11	<0.001**
10	agitatie ('agitation')	0.99	0.11	<0.001**
11	geirriteerd ('irritated')	0.99	0.10	0.001**
12	separeer ('seclusion room')	0.99	0.10	<0.001**
13	loopt ('walks')	0.99	0.11	0.015*
14	grond ('ground')	0.98	0.10	<0.001**
15	aanvang ('commencement')	0.98	0.11	0.012*
16	mede ('also')	0.98	0.10	0.001**
17	dhr wilde ('Mr wanted')	0.98	0.10	0.001**
18	liep ('walked')	0.98	0.10	0.006**
19	geagiteerd ('agitated')	0.96	0.10	0.010*
20	cvd (n/a)	0.96	0.10	0.004**

no assessment method has shown both high sensitivity and high specificity, characterizing both the difficulty of performing violence risk assessment and the need for further improvements.

VRA is a research topic that has been thoroughly described, and predictive validity of many existing methods, such as VRA checklists and unstructured clinical judgment, has been reported in literature. Although our study, based on other datasets, does not allow making strong claims about whether machine learning improves predictive validity (Kattan, 2011), we note that our internally validated predictive validities of AUC = 0.80 and 0.76 lie in the same range of existing methods—while overcoming some of their drawbacks. For example, a study by Fazel et al. (2012) assessed median predictive performance of the four most commonly used VRA instruments over 30 different studies (AUC = 0.72 [IQR 0.68–0.78]), while another study by Teo et al. (2012) assessed the level of accuracy of psychiatric residents (AUC = 0.52) and trained psychiatrists (AUC = 0.70). A study by Suchting et al. (2018) performed automatic VRA based on roughly 300 structured variables with comparable performance to our approach (AUC = 0.78).

The terms obtained in exploratory analysis, before application of modeling techniques, demonstrate a potential new type of risk factors that should be taken into account. VRA instruments are often based on a combination of static factors (e.g. previous violent behavior, employment status) and dynamic factors (e.g. hostility, disorder symptoms). The terms we extracted from text are mostly dynamic, and pertain to behavioral cues (e.g. *angry*, *walked*) and social interactions (e.g. *reacts*, *offered*), which may be more difficult to capture in a structured instrument but appear to provide important additional information.

A major strength of our research is the translational value that is obtained by using clinical notes from the EHR. Clinical text is already recorded as part of treatment by a majority of psychiatric health care institutions, implying that our machine learning approach can be widely used to support violence management in daily practice. Second, applying a flexible machine learning approach allows method customization to local requirements, and furthermore reveals the predictive validity for the relevant population—especially of importance given the lack of robustness and generalizability of existing models and methods. Finally, much attention has been devoted to the *actuarial* vs *clinical* debate (Monahan and Skeem, 2014), pertaining to the question whether actuarial VRA instruments, or VRA instruments based on clinical judgment are superior. Our approach essentially combines both approaches, by using clinical judgment captured in clinical notes as input for an actuarial tool. This allows leveraging of practitioners clinical experience, while establishing a reasonably

objective judgment through subsequent statistical modeling.

One limitation of our study is the fact that the data obtained from EHRs is originally designated for treatment rather than research. This introduces some noise to our dataset, both in clinical notes and in violence incident reports, for example in reporting discrepancies among different wards. This source of measurement uncertainty cannot be quantified, warranting some caution when interpreting our results. Furthermore, we predominantly used AUC, a measure of discrimination, to measure the predictive validity of our models. This measure is known to have some limitations, such as inability to account for prevalence (Halligan et al., 2015). Finally, we used a black box modeling approach combining the paragraph2vec and Support Vector Machine algorithms to assess violence risk, inhibiting a straightforward substantiation of probability of violent behavior. Although the terms obtained in exploratory analysis, together with the subgroup analysis of predictive validity, have elucidated the problem context to some extent, they do not directly explain model behavior. How such explanations can reliably be obtained, both at the patient and at the model level, is still a topic of ongoing research in computer science (Miller, 2019). An exploration of model explainability is included in Supplementary Materials (Section 6.5.5 and Figure 6.2).

Before an automatic VRA approach can be used in practice, some important challenges need to be addressed. Our results point out that both high sensitivity and specificity are unlikely to be achieved simultaneously. Further research is needed to point out the desired balance between false positives and false negatives, and hence, whether our prognostic models are most useful to identify patients at high or at low risk of violence. Additionally, what level of substantiation is necessary before automatic violence risk assessment can be used in practice also remains an open question, that should be addressed in discussion with professionals in the field.

In the near future, we envision that further advancements towards a data-driven psychiatric practice will be made, and that EHR data becomes an even more valuable asset in supporting important decisions in the clinical process. Machine learning approaches have been able to contribute substantially in other fields of medicine, and our study provides evidence that such progress is possible in mental health care as well. Although some crucial challenges need to be addressed before adoption is possible, our study highlights the potential value of EHR data, and clinical notes in particular, for decision support. Such support systems may in the future be widely applied in daily practice, contributing to more effective and efficient psychiatric treatment.

6.5 Supplementary Materials

6.5.1 Dataset Formation

Our study makes extensive use of data that is routinely collected in patients' Electronic Health Records (EHRs). A major advantage of this data is that it has already been collected for the sake of delivering quality treatment, and can be made available for retrospective research designs in many cases. This advantage is simultaneously its major weakness: since this data lacks a study protocol to guarantee adequate measurements, important data quality challenges need to be addressed. High effort is associated with unlocking this type of data for research. Before any modeling can commence, several issues need to be resolved, for example regarding data extraction (i.e. obtaining relevant data entities from EHR systems), data provenance (i.e. tracking data points to their origin in the EHR), and data preparation (i.e. applying appropriate transformations to data). If the goal of a study using retrospective EHR data is to obtain new insights with confidence, serious attention needs to be devoted to these steps.

In the Department of Psychiatry of the University Medical Center Utrecht (site 1), we initiated the Psydata Project in 2015, with the goal to improve care in daily practice by obtaining new insights and decision support through analysis of retrospective EHR data. After proving feasibility of such a project (Menger et al., 2016), we set out to structurally address challenges such as mentioned above (data extraction, data provenance, data preparation) in order to learn from EHR data. Since this is a relatively new area of research, we solved this problem by developing the Capable Reuse of EHR data (CARED) framework (Menger et al., 2019a). We identified the most important challenges of reusing EHR data and then proposed the framework for infrastructure that can address them. A technical infrastructure based on this framework was implemented, to support our EHR data analysis goals. This technical infrastructure for example addresses reproducibility of research, data preparation, collaboration among researchers, and documentation of code and data. Together with organizational artefacts such as guidelines for documentation and internal control, this can guarantee data quality. Our current practices ensure an up-to-date, de-identified, and accessible dataset of most information that is recorded in the EHR. The system is maintained by a multidisciplinary team of professionals, that documented the process and data in detail. The dataset used for this study, consisting of information about admissions, incidents, and clinical notes, is thus a result of careful deliberation among both data analysts and practitioners, securing its validity.

At Antes (site 2), EHR data are extracted to a clinical data warehouse that is designed largely in line with the requirements of CARED. Given the goal of attempting to replicate findings in site 1, for defining the cohort and selecting the data we followed choices that were mandated by the study design in site 1. Where knowledge specific to this site was involved (e.g. in selecting the appropriate wards), extra attention was devoted to consult with local experts. Choices in both sites were finally discussed in a focus group with stakeholders from both sites present, in order to check whether any discrepancies between choices in both sites existed. No such discrepancies were identified during the meeting, guaranteeing a similar dataset with the same standard for data quality.

6.5.2 Paragraph2vec Model Training

Since classification models use numbers as input rather than text, a suitable vector representation of clinical notes is needed before classification can occur. For this purpose we have used the paragraph2vec algorithm (Le and Mikolov, 2014), which is an extension of the earlier word2vec algorithm (Mikolov et al., 2013). Both algorithms operate on the principle of learning a vector representation of arbitrary dimensionality using a large corpus of relevant text. This is achieved by training a neural model with a hidden layer to predict target words (i.e. a word in a sentence) based on its context words (i.e. its surrounding words). The learning process takes place in an unsupervised way, meaning that no outcome variable or document labels are needed to learn accurate vector representations. The word2vec algorithm produces a corresponding vector in the vector space for each word in the training corpus. Its main advantage over a simple bag-of-words approach is that word2vec vector representations allow vector operations such as addition, subtraction, and cosine similarity, that can produce semantically meaningful results. The paragraph2vec algorithm produces a corresponding vector for each document in the training corpus, and additionally also allows inferring vectors for unseen texts. This is a probabilistic process, that works by fixing the weights of the neural model and optimizing a randomly initialized representation vector, rather than the other way round during representation training.

Since clinical text is a domain-specific language that can contain idiosyncrasies, spelling errors, and terms that have domain-specific meanings, pretrained paragraph2vec models that are for instance trained on Wikipedia or Google News data do not necessarily yield useful representations for clinical notes. For this reason, in both sites we obtained a large internal set of de-identified clinical notes, both with at least 1 million notes, to train para-

Table 6.5: Chosen paragraph2vec model settings.

Parameter	**Value**
Batch size	10,000
Epochs	20
Learning rate	0.025–0.001
Learning rate decay	Linear
Minimum count	20
Model	Distributed Memory
Sub sampling	0.001
Vector size	300
Window size	2

graph2vec models. As preprocessing steps, we transformed all text to lowercase, remapped special characters, and removed all characters that were not whitespace, period, or alphabetical characters. We then tokenized text (i.e. split it to words), removed stop words, and applied stemming (i.e. mapping inflections of words to their stem). The resulting sequence of terms was then used to train a paragraph2vec model. Optimal paragraph2vec model settings are still a topic of ongoing research, we based our choices on default model settings in the Gensim package (Řehůřek and Sojka, 2010) that was used for training, in combination with information by Chiu et al. (2016) and Lau and Baldwin (2016) (Table 6.5). We used the Distributed Memory model for training the algorithm, which concatenates input vectors, and is thus able to take word order into account. Model dimensionality typically ranges between 100 and 1000, we opted for a dimensionality of 300 as a middle ground. We slightly decreased the window size from 5 to 2, and increased the minimum word count from 5 to 20 to mitigate effects of lack of structure and spelling errors present in clinical text. We increased the number of epochs to 20, in order to increase the likeliness of reaching model convergence on our data set. Other parameters were not changed from Gensim defaults. The result of training includes two independent paragraph2vec models that comprise the machine learning pipeline together with the classification models.

In order to determine numerical representations of clinical notes in our dataset using the trained paragraph2vec model, we first concatenated all relevant notes for a single admission, and then averaged over ten paragraph2vec inferences of this unseen concatenation of notes, to cancel out inaccuracies due to the probabilistic nature of the inference.

6.5.3 Cross Validation Procedure

When applying machine learning models to a dataset, one must ensure that data is never simultaneously used to train and test a model. Information leakage between these two sets will inevitably lead to overly optimistic estimates of model predictive validity. We chose a nested cross validation procedure, to simultaneously optimize, train, and assess the predictive validity of a model on a single dataset while obtaining a reliable estimate of performance without bias.

Our classification model consists of a Support Vector Machine with a radial kernel. This type of machine learning algorithm has two hyperparameters that should be optimized: the cost parameter (C) that determines how strong models during training are penalized for data points on the wrong side of the classification boundary, and the gamma (γ) parameter that determines how far the influence of a single training example reaches. We determined the optimal values for these parameters using a grid search, i.e. by training a Support Vector Machine for multiple combinations of C and γ values. For C we chose a range of $[10^{-1}, 10^{0}, 10^{1}]$, and for γ we chose $[10^{-6}, 10^{-5}, \ldots, 10^{0}]$. We chose a relatively narrow range for C because models trained on our dataset are empirically not very sensitive to this parameter, and to speed up model training time. Model performance was then estimated on a hold-out set, i.e. a subset of data that is not used for training models. Since using one single hold-out set can introduce bias into performance estimates, we used cross validation to repeat this process five times on non-overlapping test sets, and chose hyperparameters that perform best on average. This procedure comprises the inner cross validation loop $\mathrm{CV}_{\mathrm{inner}}$.

A new model was then trained using data of all five $\mathrm{CV}_{\mathrm{inner}}$ folds, using the optimal hyperparameters found in the $\mathrm{CV}_{\mathrm{inner}}$ loop, and performance was tested on yet another hold-out set. For the same reasons as mentioned above, we repeated this procedure in five folds as well, in the $\mathrm{CV}_{\mathrm{outer}}$ cross validation. While the $\mathrm{CV}_{\mathrm{inner}}$ loop is used to determine optimal hyperparameters, the $\mathrm{CV}_{\mathrm{outer}}$ loop is used obtain a reliable estimate of performance on unseen data. In both cross validation loops, we furthermore ensured that data points from the same patients (i.e. previous or future admissions) were always grouped in the same fold, mainly to prevent information of future admissions influencing performance assessment.

Given the five folds $\mathrm{test}^1_{\mathrm{outer}}, \ldots, \mathrm{test}^5_{\mathrm{outer}}$ that were used to estimate performance in $\mathrm{CV}_{\mathrm{outer}}$, we computed the Area Under Curve by averaging over the five folds, i.e. using $\mathrm{AUC} = \frac{1}{5}\sum_{i=1}^{5} \mathrm{AUC}(\mathrm{test}^i_{\mathrm{outer}})$. To estimate standard error of the mean of AUC, we used the DeLong method (DeLong et al., 1988) for

estimating variance of AUC for each fold using $\text{VAR}^i = \text{delong-var}(\text{test}^i_{\text{outer}})$. The DeLong method is applicable in this case, and preferred when other methods based on bootstrapping are computationally not feasible (Hajian-Tilaki and Hanley, 2002). We then computed average variance (VAR) over the five folds $\text{VAR} = \frac{1}{5}\sum_{i=1}^{5}\text{VAR}^i$, and took the square root to compute the average standard deviation $\text{SD} = \sqrt{\text{VAR}}$. To estimate the standard error of the mean AUC we finally used $\text{SE} = \text{SD}/\sqrt{5}$, given AUC samples in five different folds. Other outcome statistics were determined based on a 2x2 contingency table, showing true negatives, false negatives, false positives, and true positives. To map classification probabilities (i.e. probability of showing violent behavior) to a binary outcome, we set a classification threshold so that classification has the same distribution as outcome (i.e. the true labels). This ensures false positives and false negatives are balanced, as the optimal balance for daily practice still needs to be established. The classification threshold was set per fold, because predictions among different folds are not necessarily calibrated with regard to each other. The contingency table was finally determined by summing per-fold contingency tables, and other statistics such as sensitivity and specificity are determined based on this contingency table.

Results of the hyperparameter optimization procedure are displayed in Table 6.6. The Area Under Curve (AUC) values are based on the internal cross validation loop. These values are relatively close to outer cross validation results, showing that model convergence has been reached, while the cross validation setup has inhibited overtraining of models.

6.5.4 Code and Data Availability

We have made all analysis code, predictions, and output logs available in an online GitHub repository. This allows other researchers to verify that analysis was indeed performed as described in this paper, makes data accessible for meta studies, and allows potential reproduction of our results by other researchers on new and independent datasets.

The analysis code is available in the form of a set of Jupyter Notebooks, which implement Python scripts developed for this study. Predictions and output logs are available in csv and text files. The GitHub repository is accessible online, where additional information and documentation can be found.[1]

All data associated with the study is stored internally, so that results can be verified, and modifications of method or potentially new modeling techniques

[1] `http://www.github.com/vmenger/violence-risk-assessment`

can be applied in the future. Unfortunately, the dataset cannot be shared with other researchers due to legal and privacy constraints.

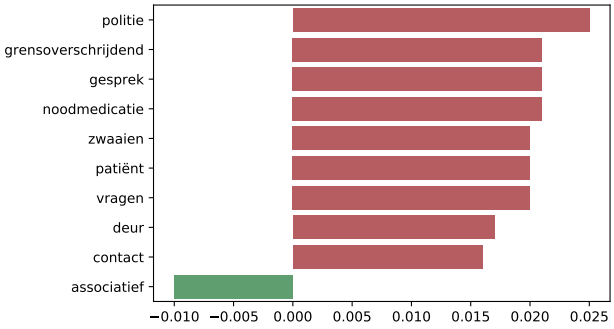
6.5.5 Model Explainability

In order to explore whether classification model behavior at the local level can be explained, we applied the Linear Model-Agnostic Explanations (LIME) method (Ribeiro et al., 2016) to our trained models. This method tries to approximate the decision boundary near a specific data point using a linear function. Specifically, it samples points around a data point that is to be explained, and uses the trained machine learning pipeline to classify this set of data points. Based on these data points and their classified outcome, LIME trains a k-lasso model on a bag-of-words representation of sampled data points, returning a set of k terms in these texts that are relevant for the local decision boundary.

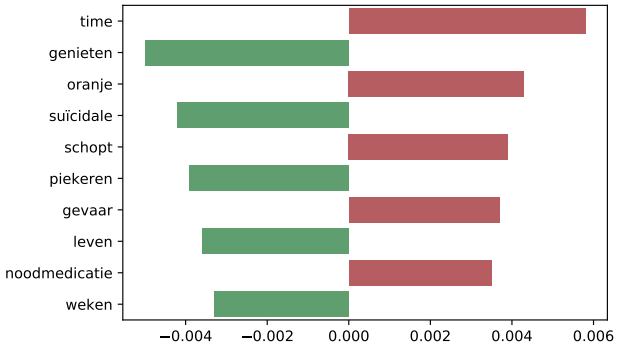
Based on an exploratory evaluation that presented explanations of a small subset of data points to eight human subjects, we found that presenting an explanation (e.g. Figure 6.2) in combination with a risk assessment increased participants' trust in the system. We additionally found no evidence of bias (e.g. discrimination against protected groups) in the classification model. Some points of model failure were finally identified, where texts were classified using, apparently to the human user, arbitrary terms. This information can be used as feedback to improve the dataset and trained models.

Table 6.6: Optimal hyperparameters and optimal AUC based on the inner cross validation loop, where a grid search is performed to determine the optimal Support Vector Machine parameters. The Inner AUC column shows the average AUC over five inner folds, while the Outer AUC column specifies performance on the hold-out fold for this iteration. Abbreviations: C = cost, γ = inverse of kernel radius, AUC = Area Under Curve, SD = Standard Deviation.

	Site 1				**Site 2**			
Fold	C	γ	**Inner AUC (SD)**	**Outer AUC**	C	γ	**Inner AUC (SD)**	**Outer AUC**
1	0.1	0.001	0.797 (0.008)	0.755	1.0	0.01	0.753 (0.044)	0.761
2	0.1	0.001	0.792 (0.038)	0.813	1.0	0.01	0.759 (0.029)	0.741
3	0.1	0.001	0.784 (0.019)	0.805	10.0	0.01	0.742 (0.044)	0.784
4	1.0	0.0001	0.790 (0.038)	0.792	1.0	0.01	0.741 (0.039)	0.782
5	0.1	0.001	0.786 (0.040)	0.818	1.0	0.01	0.762 (0.038)	0.752



(a) Example 1



(b) Example 2

Figure 6.2: Two examples of local explanations of models. The explanation on top (a) predicted high risk of aggression, which is reflected in terms such as *politie* ('*police*') and *noodmedicatie* ('*emergency medication*'). The explanation on the bottom (b) predicted low risk, explained by terms such as *genieten* ('*enjoy*') and suïcidale ('*suicidal*'), but also exhibits high-risk terms such as *schopt* ('kicks') and *gevaar* ('danger').

Table 6.7: For both the internal and external analysis, and for both sites, a subgroup analysis pertaining to early vs. late violence. Abbreviations: AUC = Area Under Curve, CI = Confidence Interval, CV = Cross Validation.

Evaluation	**Day of first violence incident (median)**	**Early violence AUC (95% CI)**	**Late violence AUC (95% CI)**	**Difference (95% CI)**	***p*-value**
Internal (CV), site 1	6	0.821 (0.787 to 0.854)	0.775 (0.740 to 0.810)	0.046 (-0.003 to 0.094)	0.06
Internal (CV), site 2	9	0.771 (0.724 to 0.818)	0.755 (0.708 to 0.803)	0.015 (-0.051 to 0.082)	0.65
External model, site 1	6	0.745 (0.704 to 0.785)	0.698 (0.652 to 0.744)	0.046 (-0.015 to 0.108)	0.14
External model, site 2	9	0.653 (0.609 to 0.698)	0.632 (0.587 to 0.678)	0.021 (-0.042 to 0.085)	0.51

Table 6.8: For both the internal and external analysis, and for both sites, a subgroup analysis pertaining to short vs. long admissions. Abbreviations: AUC = Area Under Curve, CI = Confidence Interval, CV = Cross Validation.

Evaluation	**Length of admission (median)**	**Short admissions AUC (95% CI)**	**Long admissions AUC (95% CI)**	**Difference (95% CI)**	***p*-value**
Internal (CV), site 1	16	0.805 (0.764 to 0.846)	0.792 (0.758 to 0.826)	0.012 (-0.041 to 0.066)	0.65
Internal (CV), site 2	15	0.789 (0.738 to 0.839)	0.730 (0.686 to 0.774)	0.058 (-0.008 to 0.125)	0.09
External model, site 1	16	0.704 (0.653 to 0.755)	0.736 (0.700 to 0.775)	0.032 (-0.033 to 0.096)	0.34
External model, site 2	15	0.655 (0.607 to 0.702)	0.633 (0.589 to 0.678)	0.022 (-0.043 to 0.087)	0.52

7 | Using Cluster Ensembles to Identify Psychiatric Patient Subgroups

Identification of patient subgroups is an important process for supporting clinical care in many medical specialties. In psychiatry, patient stratification is mainly done using a psychiatric diagnosis following the Diagnostic and Statistical Manual of Mental Disorders (DSM). Diagnostic categories in the DSM are however heterogeneous, and many symptoms cut across several diagnoses, leading to criticism of this approach. Data-driven approaches using clustering algorithms have recently been proposed, but have suffered from subjectivity in choosing a number of clusters and a clustering algorithm. We therefore propose to apply cluster ensemble techniques to the problem of identifying subgroups of psychiatric patients, which have previously been shown to overcome drawbacks of individual clustering algorithms. We first introduce a process guide for modeling and evaluating cluster ensembles in the form of a Meta Algorithmic Model. Then, we apply cluster ensembles to a novel cross-diagnostic dataset from the Department of Psychiatry of the University Medical Center Utrecht in the Netherlands. We finally describe the clusters that are identified, and their relations to several clinically relevant variables.

This work was originally published as:
Menger, V., Spruit, M., van der Klift, W., and Scheepers, F. (2019b). Using cluster ensembles to identify psychiatric patient subgroups. In *Artificial Intelligence in Medicine*, pages 252–262. Springer International Publishing

7.1 Introduction

Identification of patient subgroups is an important process that is able to guide clinical treatment in many medical specialties. In psychiatry, the main construct for stratifying patients is a psychiatric diagnosis, typically performed using the Diagnostic and Statistical Manual of Mental Disorders (DSM). This manual describes various high level disorders such as depressive disorders, anxiety disorders, and developmental disorders, with sub-types for each category. It defines clear diagnostic criteria based on symptoms—a major depressive disorder for instance can only be diagnosed after eight symptoms have been assessed, including depressed mood, weight loss, fatigue, and inability to concentrate, and at least five were observed in a two-week period. While the DSM is by far the most widely adopted standard for diagnosis, in recent years its rigid approach has been subject to criticism. Research for instance shows that the DSM has little biological validity (i.e. lack of connection to biomarkers), that diagnostic categories are not specific (i.e. large heterogeneity exists within groups), and that symptoms often cut across diagnostic categories (Cuthbert and Insel, 2013).

This critique on the DSM has seeded data-driven approaches that seek interesting subgroups using relevant datasets rather than using expert elicited criteria. For this purpose, various clustering algorithms that are able to discover latent subgroups have been applied to patient data. One major downside of a clustering approach however is the need to select an appropriate number of clusters and an appropriate clustering algorithm, which both have been shown to provide challenges for researchers (Kuncheva and Hadjitodorov, 2004). The majority of studies rely on a single metric for choosing the right number of clusters, and subsequently apply a single clustering algorithm (Marquand et al., 2016), while both choices can have significant impact on the results that are obtained. Consequently, as of yet no consensus exists on either the number or nature of psychiatric patient subgroups that can be derived in this data-driven way.

In this work we therefore propose to apply cluster ensembles, i.e. combinations of multiple clustering algorithms, to this problem. This enables identification of distinct subgroups that can directly inform treatment, while overcoming the downsides of individual clustering algorithms. Previous work has already shown that cluster ensembles often improve robustness, stability, and accuracy over individual clustering algorithms, both in general and in the medical domain, yet this approach is still rare in mental health care research (Topchy et al., 2003; Ghaemi et al., 2009).

The contribution of this work is twofold. First, we present a process guide for modeling and evaluating cluster ensembles in the form of a Meta Algorithmic Model, as introduced in (Spruit and Jagesar, 2016). This guide aims to support researchers in applying cluster ensembles in their particular (medical) domain. Second, we apply a cluster ensemble approach to a novel cross-diagnostic dataset of 1,098 Youth Self Report (YSR) questionnaires of adolescents that were treated at the Department of Psychiatry of the University Medical Center Utrecht (UMCU) in the Netherlands. Since these questionnaires were routinely captured during treatment, using them to identify patient subgroups, if present, can have direct applicability in the psychiatric practice (Menger et al., 2016, 2019a). After applying the cluster ensemble approach, we examine key characteristics of the clusters we obtained, and assess their relation to several clinically relevant variables including DSM diagnosis.

7.2 Background and Related Work

Clustering algorithms have previously been used in mental health care research for stratifying patients with a common psychiatric diagnosis, such as schizophrenia, depression, or autism (Marquand et al., 2016). The number of clusters ranges from two to seven, typically selected based on a single measure such as Bayesian Information Criterion or Ward's method. Most researchers then apply one algorithm to their dataset, such as K-means Clustering, Hierarchical Clustering, or Latent Class Cluster Analysis. Clusters of various natures have been found, for instance based on differences in symptoms (Dollfus et al., 1996), treatment outcome and onset (Cole et al., 2012), and patient functioning (Fountain et al., 2012). A smaller number of studies focused on stratifying patients in a cross-diagnostic setting. A study by Olino et al. (2010) for instance found six subtypes differing in presence of depression, anxiety, or a mixture of both, while Lewandowski et al. (2014) reported a neuropsychologically normal subtype, a globally impaired subtype, and two mixed cognitive profiles, and Kleinman et al. (2014) found a cluster with diminished sustained attention, inhibitory control and vigilance, and increased impulsiveness, and a second converse cluster. So far, cluster ensembles have only been applied once in mental health research in a study by Shen et al. (2007) who used this technique to identify four subtypes of pervasive developmental disorders. They reported differences in severity, in problems with language acquisition and impairment, and in aggressive behavior.

To reduce variability in clustering outcomes, such as for example described above, cluster ensembles were proposed based on the principle that multi-

ple weak partitions in combination can provide a more accurate and objective outcome than a single strongly optimized clustering (Fred, 2001). This is analogous to ensemble learning techniques such as Boosting and Random Forests in the supervised domain. First, during the generation stage, a number of diverse partitions are obtained, ideally with strengths and weaknesses in different parts of the solution space (Ghaemi et al., 2009). This is for instance achieved by using multiple clustering algorithms and different algorithm parameters, by subsetting data, and by projecting data to subspaces (Topchy et al., 2003). The result of the generation stage on a dataset $X = \{x_i, \ldots, x_n\}$ with n observations is a partition set $P = \{p_1, \ldots, p_m\}$ of m partitions, where each $p_i = \{C_1^i, \ldots, C_k^i\}$ assigns every observation to a single cluster C_i out of k clusters. In the subsequent consensus stage, an optimal partitioning is obtained using partition set P. Various procedures have been proposed based on object co-occurrence in clusters, such as majority voting (Topchy et al., 2004), or the graph-based Cluster-based Similarity Partitioning Algorithm (CSPA) (Strehl and Ghosh, 2003). Another type of approach finds the median partition in $p^* \in P$, for instance defined as the partition that maximizes similarity with all other partitions in P (Vega-Pons and Ruiz-Shulcloper, 2011). Cluster ensembles have recently successfully been applied in several biomedical domains (Boongoen and Iam-On, 2018).

7.3 Meta Algorithmic Model

To support researchers in applying cluster ensembles to their (medical) domain, we propose a Meta Algorithmic Model (MAM) of cluster ensemble modeling and evaluation (Figure 7.1). Our MAM is an extension of the original work of Spruit and Jagesar (2016), that was aimed at supervised learning tasks. In their words, MAMs are intended to provide "highly understandable and deterministic method fragments — i.e. activity recipes — to guide application domain experts without in-depth Machine Learning expertise". Method fragments are specified as a combination of a Unified Modeling Language (UML) activity diagram showing processes, and a UML class diagram showing concepts.

The cluster ensemble modeling process, shown on the left of Figure 7.1, starts with loading a prepared dataset. Then, in the generation stage multiple methods for introducing diversity in the cluster portfolio are used, including observation and feature sampling, choosing clustering algorithms and selecting a number of repetitions. After a number of clusters and a distance measure are selected, the cluster portfolio is created. In the subsequent consensus stage

a consensus function should be selected, and weak partitions can be trimmed from the cluster portfolio. During the evaluation stage, internal index criteria (e.g. Calinski-Harabasz, Silhouette) can be evaluated, and clusters can be visualized after applying a dimension reduction algorithm to the dataset. Cluster characteristics can be identified based on the cluster assignments of the dataset, and an external evaluation (e.g. using expert evaluation, or comparison to a reference class) can finally be performed. The class diagram on the right of Figure 7.1 shows which concepts need to be instantiated in relation to each process step.

7.4 Applying Cluster Ensembles

7.4.1 Dataset

We applied the cluster ensemble modeling approach in Figure 7.1 to a novel cross-diagnostic dataset of adolescent patients who were treated at the Department of Psychiatry of the UMCU. The dataset consisted of Youth Self Report (YSR) questionnaires, a standardized checklist aimed at adolescents. It consists of 112 items in the form 'I am/have/feel *symptom/behavior*', which a respondent can indicate as 'not true', 'somewhat or sometimes true', and 'very true or often true'. The YSR defines eight outcome scales by summing responses of specific item subsets: (1) Anxious depressed, (2) Withdrawn depressed, (3) Somatic complaints, (4) Social problems, (5) Thought problems, (6) Attention problems, (7) Rule breaking behavior, and (8) Aggressive behavior. We dismissed 50 reports with more than five percent out of 112 items missing, and for the remaining YSRs imputed missing values with the median score of that item. If multiple reports of a patient were present ($n = 175$), we used only the first report, under the assumption that treatment effect is smallest at this point. Our final dataset consists of 1,098 YSRs. The mean age of respondents was 14.7 years (SD = 2.2), and 44.5% of respondents were female.

For cluster ensemble modeling, we used the eight outcome scales of the YSR as input data. Since these scales have a non-arbitrary zero value (i.e. absence of any symptoms), we chose to analyze them as ratio scales, using Euclidean distance, implicitly assuming equidistant item scores. Since the outcome scales are a sum of individual items measured on an ordinal scale, they could also be regarded as ordinal scales themselves. However, this distinction is often relatively unimportant in practice, especially when performing clustering (Harpe, 2015). Analyzing these data as ratio scales furthermore allows a larger variety

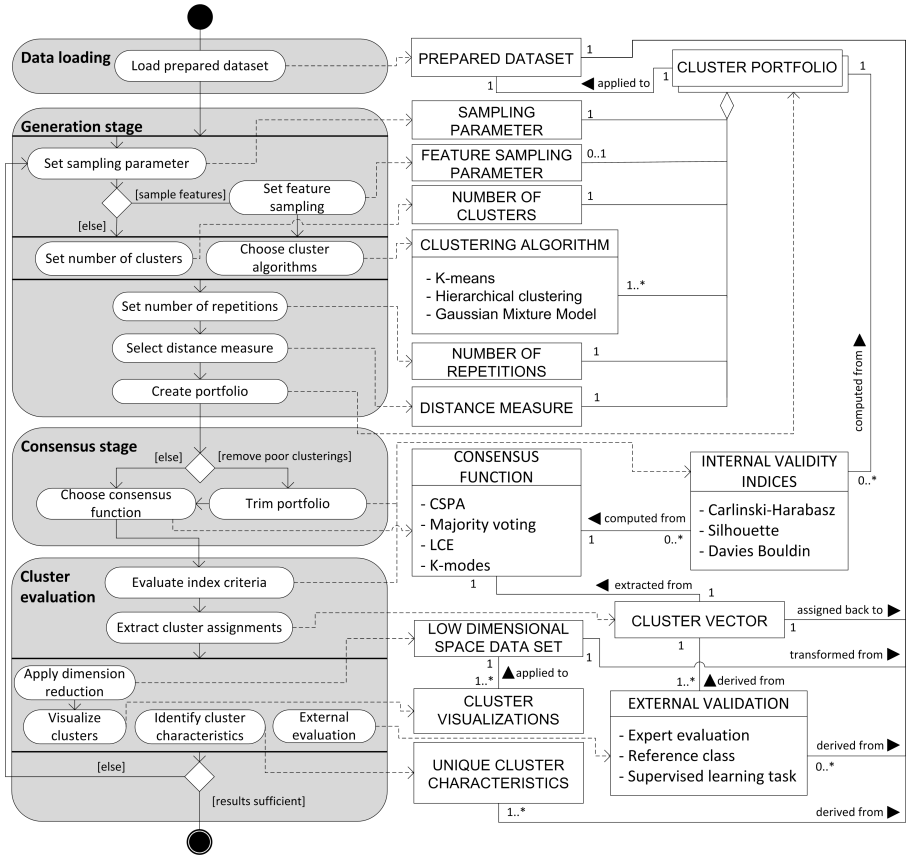


Figure 7.1: Method fragment of the Meta Algorithmic Model for cluster ensemble modeling and evaluation. Abbreviations: CSPA = Cluster-based Similarity Partitioning Algorithm, LCE = Link-Based Cluster Ensemble.

of clustering algorithms to be applied to this dataset, most likely improving clustering outcomes.

7.4.2 Cluster Ensemble Modeling

One risk of performing cluster analysis is obtaining clusters, while no natural grouping exists in a dataset. For this purpose, we computed the Hopkins statistic as a measure of clustering tendency (Banerjee and Dave, 2004). This statistic is computed from a dataset X with n observations by creating a sample $Y \subseteq X$, and a set of uniform randomly sampled points U, with U and Y both of size $m \ll n$. Then, let q_i be the distance of $u_i \in U$ to its nearest neighbor in X, and let p_i be the distance of $y_i \in Y$ to its nearest neighbor in X, according to some distance measure d. The Hopkins statistic is finally given by:

$$H = \frac{\sum_{i=1}^{m} q_i}{\sum_{i=1}^{m} p_i + \sum_{i=1}^{m} q_i} \tag{7.1}$$

The Hopkins statistic ranges from 0 (uniformly distributed data), to 0.5 (randomly distributed data) to 1 (highly clusterable data). Computing this statistic for our dataset using Euclidean distance obtains $H = 0.71$. No definitive cut-off for cluster tendency has been established, but a value between 0.5 and 1 is regarded as indicative for high likelihood of significant clusters.

Next, an appropriate number of clusters k should be selected. Rather than rely on a single measure for determining this number, we used the R package `NbClust`, which computes 26 internal validity indices for several values of k, and proposes an optimal number of clusters based on a majority vote. We computed the validity indices in combination with both K-means clustering and hierarchical clustering, and set the number of clusters between two and seven. The majority vote shows that the optimal number of clusters $k = 3$ for our dataset, which we will use in all following steps.

For application of the cluster ensemble to our dataset, the R package `DiceR` was used, which implements various cluster ensemble techniques. In order to find an appropriate subset of algorithms, we applied each of the twelve implemented algorithms with their standard settings to the dataset. We then selected three algorithms that obtain different partitions of the dataset, based on two-dimensional Principal Component Analysis (PCA) plots. These are the K-means algorithm (Figure 7.2a), which minimizes the within-cluster sum of squares using an iterative approach, a Gaussian Mixture Model (Figure 7.2b), which models the dataset with a mixture of multi-dimensional Gaussian probability distributions, and the Affinity Propagation algorithm (Figure 7.2c),

which approaches a dataset as a network in which data points communicate with all other points.

To obtain a diverse cluster portfolio, we used five reruns for each of the three clustering algorithms with a random subset of 80% of all data. The number of clusters k is fixed to three, as determined previously. We trimmed the cluster portfolio using a Rank Aggregation method: all partitions were ranked based on several internal validity indices, and the 75% highest partitions were retained. We finally used the Cluster-based Similarity Partitioning Algorithm (CSPA) to obtain a single clustering based on the cluster portfolio (Figure 7.2d). All analysis code is made publicly available on GitHub.[1]

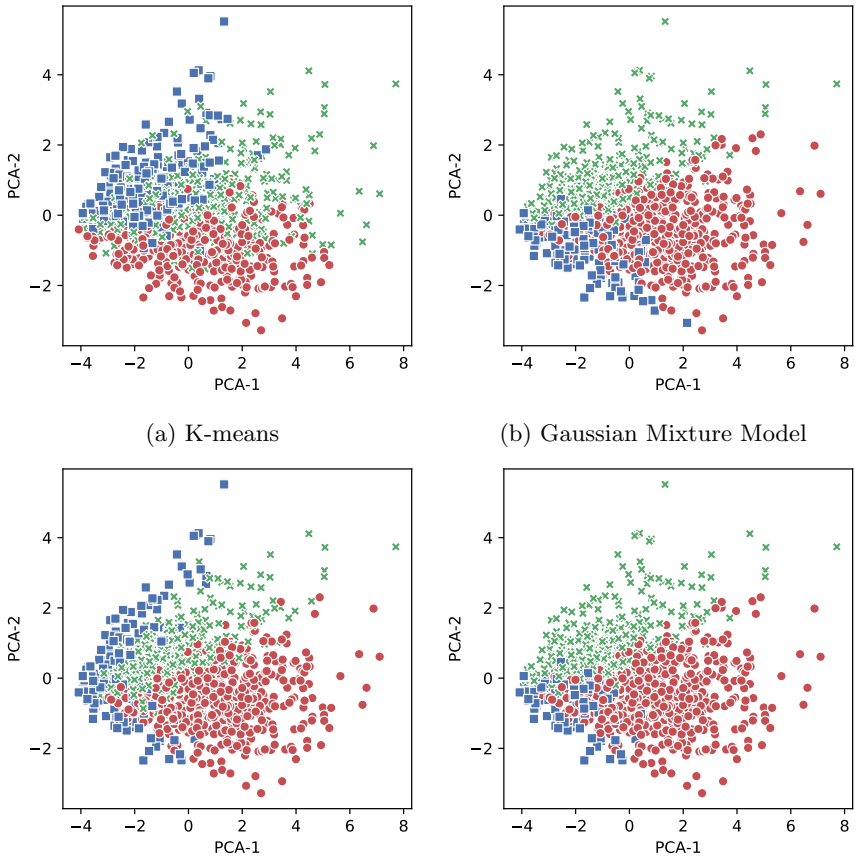
7.5 Cluster Evaluation

Applying the cluster ensemble method to our dataset results in three clusters, which contain respectively 55.5%, 32.1%, and 12.5% of observations (Figure 7.2d). The ensemble clustering shows strongest similarity with the Gaussian Mixture model, with some differences in the two smallest clusters, and greater differences with the K-means and Affinity Propagation partitions. To assess statistical significance of the three clusters found by the cluster ensemble approach, we used the sigClust method (Huang et al., 2015) which tests against a null hypothesis of all data being from a single Gaussian distribution. This results in $p = 0.01$ when applied to our dataset, indicating presence of significant clusters at the $\alpha = 0.05$ level.

Figure 7.3 shows the median value of the eight YSR scales over the three clusters, where distinctions among the three clusters can be observed. Cluster 1, the largest cluster, has the highest overall scores, especially in the two depressed scales (1–2). Values of other scales are among the highest as well in Cluster 1, with Rule Breaking Behavior being the lowest item. Clusters 2 and 3 on the other hand generally have lower scores, with equal median outcomes on the Anxious depressed, Withdrawn depressed, and Somatic problems scales (1–3). For the other five scales, Cluster 2 shows higher outcomes. For the Rule Breaking Behavior and Aggressive Behavior scales (7–8), Cluster 2 shows higher median values than Cluster 1 as well.

To identify clusters' distinguishing characteristics, we integrated clinical notes from the EHR, i.e. pieces of text written by caregivers about treatment, that were de-identified using the DEDUCE method (Menger et al., 2018b), in the two weeks surrounding YSR response. We extracted the 1,000 most

[1] `http://www.github.com/vmenger/cluster-ensembles`



(a) K-means

(b) Gaussian Mixture Model

(c) Affinity Propagation

(d) Cluster ensemble

Figure 7.2: Partitions of the dataset after applying Principal Component Analysis, both based on single algorithm (a–c), and combined in Cluster Ensemble (d). Abbreviations: PCA = Principal Component Analysis.

frequent terms from these texts, and computed the Spearman correlation coefficient for each term and each of the three clusters vs the other two clusters. A psychiatrist then selected three informative terms among those with the highest positive correlation coefficients. For Cluster 1, the selected terms are *depressive*, *dejected*, and *suicidal*, which is in line with high scores in the two depressed scales. For Cluster 2, the terms *behavioral problems*, *adhd* (attention deficit hyperactivity disorder), and *distracted* are identified, which is in line with high scores on the Attention Problems and Aggressive Behavior scales. For Cluster 3, these terms are *speech*, *verbal*, and *individual*. Based on these terms and Figure 7.3, we describe Cluster 1 as 'depressive symptoms', and Cluster 2 as 'behavioral problems'. A comprehensive description of Cluster 3 is less evident, we therefore describe it as 'low severity'.

Table 7.1 shows the three clusters versus the main DSM diagnosis, which had been made definitive within 12 weeks of YSR response for a subset of 665 patients. The most common diagnosis for the three clusters respectively are Anxiety Disorder, Attention Deficit Disorder, and Pervasive Developmental Disorder (PDD). Diagnoses are typically present in several clusters, although they are usually most prominent in one single cluster, with the exception of PDDs.

We finally integrate several clinically relevant variables, including Global Assessment of Functioning (GAF) score at start and end of treatment, a seven-point burden of disease indicator, and length of treatment (Table 7.2). Although Cluster 1 has the highest overall YSR outcome scale scores, the GAF scores both at start and end of treatment are relatively low. The difference between these groups are assessed with a Kruskal-Wallis one-way analysis of variance test. Results show that significant differences in GAF score at start and end of treatment and in length of treatment exist at the $\alpha = 0.05$ level, but not in burden of disease. This indicates that clusters do not only differ in YSR outcome scales, but also in variables that are relevant in clinical practice.

7.6 Discussion and Conclusion

In line with previous research, our results point out that different clustering algorithms indeed obtain different partitions. Cluster ensembles are a useful method to overcome such issues. By applying our proposed cluster ensemble approach to a dataset of YSR questionnaires, we obtained three distinct patient subgroups. Patients with the same DSM diagnosis are typically represented in multiple clusters, indicating that the three clusters are to some extent a novel stratification of adolescent patients. We furthermore identified

significant differences in GAF both at start and end of treatment, and in length of treatment. Although absolute differences among clusters are modest, this shows that patient subgroups do not only differ in the YSR outcome scales.

The clustering outcomes of this study are limited by both the type of data and the specific patient population that reported it. The dataset includes eight outcome scales that are general, but may not capture all dimensions of patient well-being, and whether the clusters we obtained generalize to other populations should be the topic of further research. The main contribution of this research however lies in the cluster ensemble approach, and the process guide introduced with the Meta Algorithmic Model. Such cluster ensemble approaches are able to eliminate one source of variance in reported psychiatric patient subgroups, and can thereby in the future contribute to the identification of a more robust and objective stratification of psychiatric patients.

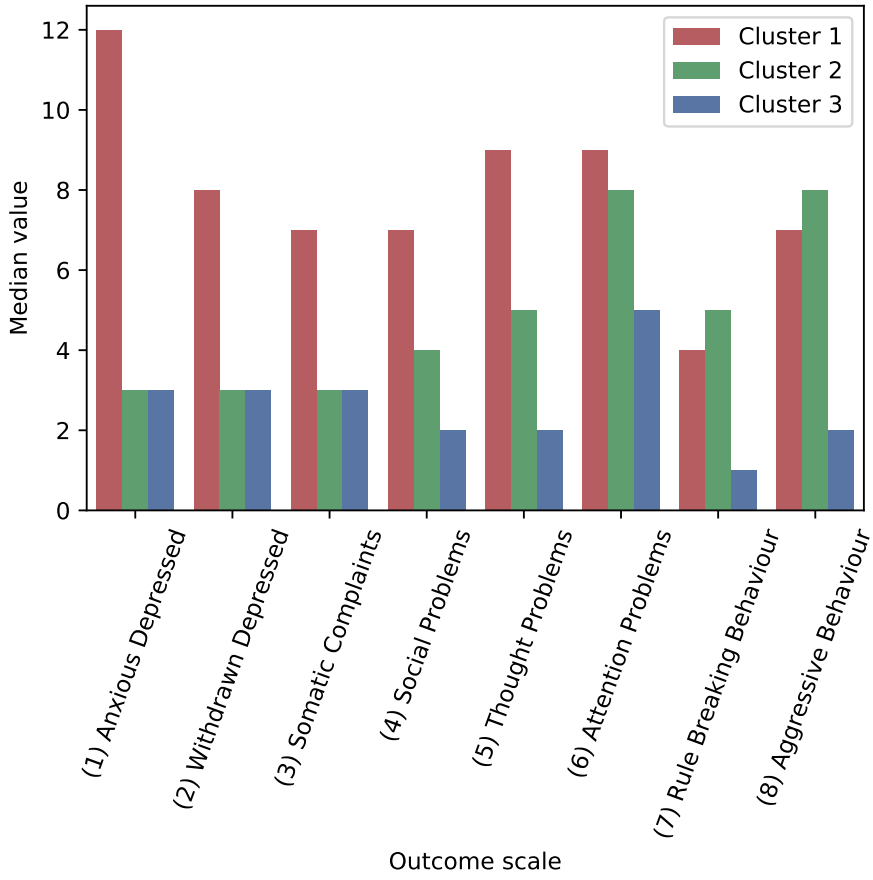


Figure 7.3: Median YSR outcome scale value for each of the three identified clusters.

Table 7.1: Main DSM diagnoses per cluster, if finalized within 12 weeks.

Disorder	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Total**
Anxiety disorder	76 (19.2%)	7 (3.5%)	13 (18.3%)	14.4%
Developmental disorder				
Attention deficit disorder	39 (9.8%)	82 (41.4%)	10 (14.1%)	19.7%
Pervasive developmental disorder	103 (26.0%)	51 (25.8%)	24 (33.8%)	26.8%
Other	15 (3.8%)	15 (7.6%)	2 (2.8%)	4.8%
Eating disorder	10 (2.5%)	2 (1.0%)	3 (4.2%)	2.3%
Mood disorder	65 (16.4%)	13 (6.6%)	7 (9.9%)	12.8%
Psychotic disorder	24 (6.1%)	7 (3.5%)	4 (5.6%)	5.3%
Personality disorder	27 (6.8%)	1 (0.5%)	0 (0.0%)	4.2%
Other	37 (9.3%)	20 (10.1%)	8 (11.3%)	9.8%
Total	396 (100%)	198 (100%)	71 (100%)	100%

Table 7.2: Average value of clinically relevant variables per cluster. p-value is assessed using a Kruskal-Wallis test, * indicates significance at the $\alpha = 0.05$ level. Abbreviations: GAF = Global Assessment of Functioning.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	p**-value**
GAF at start of treatment	45.9	50.0	46.0	0.008*
GAF at end of treatment	53.6	56.7	51.5	0.012*
Burden of disease	4.5	4.5	4.8	0.550
Length of treatment (days)	132.6	175.3	160.7	0.003*

8 | Conclusion

Motivated by the potential of knowledge discovery in Electronic Health Records to improve care, this dissertation aimed to contribute to improved knowledge discovery in psychiatric EHRs from a computing perspective. At the beginning of this research we posed the following main research question:

MRQ — How can data from Electronic Health Records provide relevant insights for psychiatric care?

In order to answer this question, the individual Chapters 2–4 of this dissertation each investigated and proposed solutions to specific technical, organizational, and ethical challenges of using EHR data for analysis. In Chapter 2, we introduced the CRISP-IDM process, in order to facilitate successful collaboration between health professionals and data experts, so that exploratory results that are supported by both can be found. In Chapter 3, we introduced the CARED framework, enabling individual health care organizations to design and implement a technical infrastructure that structurally supports reusing EHR data for analysis. In Chapter 4, we designed and implemented DEDUCE, a pattern-matching based de-identification method for Dutch clinical text—an application that was not available yet for text written in Dutch. These three chapters introduced scientifically validated artefacts that support analyzing EHR data, thereby removing some of the barriers to perform knowledge discovery in EHRs—in medicine in general, and in psychiatry in particular.

Chapters 5–7 subsequently built upon this foundation, by identifying specific problems in the psychiatric practice, and by then proposing a solution to these problems based on knowledge discovery techniques applied to EHR data. In Chapters 5 and 6, we applied machine learning techniques to the problem of violence risk assessment, and demonstrated how such an approach can assess inpatient violence risk at the start of admission with predictive validity that matches or outperforms currently employed assessment methods. In Chapter 7, we applied cluster ensembles to the problem of identifying robust subgroups of psychiatric patients, discovering three clusters of adolescent patients that differed significantly in DSM diagnosis and other clinically relevant variables. Implementing these results on the work floor—and in this way

using them to directly improve care—was not yet fully managed within the duration of this research. Both examples however have substantial potential for deployment, and show that knowledge discovery in EHRs can contribute to improving care in psychiatry.

8.1 Contributions

At the beginning of this research, we posed six research questions that aimed to investigate various aspects of the main research question. Next, we will briefly review the contributions of each individual chapter to answer these research questions, and formulate a conclusion for each of them. Together, they form an answer to the main research question.

RQ1 — How can health care professionals and data experts collaboratively perform exploratory analysis?

In order to investigate how collaboration between health care professionals and data experts is possible within an organization, we designed the Cross Industry Standard Process for Interactive Data Mining (CRISP-IDM). In this process, the Modeling and Evaluation phases of the standard process are contracted into one single iterative phase, where a selection of data is made, and additional specific data preparation is done to enable exploration of data. We used data visualization, an easy to understand way of analyzing data, as a tool to achieve this sort of exploration. Expert interviews were used to identify seven initial research themes for knowledge discovery in psychiatric EHRs: context factors, admission and dismissal, aggression, patient referrals, routine outcome monitoring, medication, and other. Subsequently, in a total of 19 sessions, 18 health care professionals collaboratively explored data from the EHR, and found 24 new hypotheses for further research that were initially not imagined. The theme which showed most promise for future research was aggression, followed by context factors, admission and dismissal, and routine outcome monitoring.

By allowing interaction and collaboration between health care professionals and data experts, we found that they were enabled to use the others' strengths in performing data analysis. Our CRISP-IDM process succeeded in facilitating such interaction, and is thus a useful approach to incorporate domain knowledge in knowledge discovery. We furthermore demonstrated how the exploratory approach is an important benefit of using EHR data for analysis. The high cost and effort associated with data collection in traditional

study designs usually inhibits such an exploratory approach. The availability of EHRs allows composing such datasets for exploratory research and generating hypotheses, that can subsequently be tested in other studies.

Conclusion I — Using the CRISP-IDM process, health care professionals and data experts can use data visualization to collaboratively find new knowledge and unexpected hypotheses. This approach proves mainly useful at the start of a data analysis project, so that health care professionals and data experts can familiarize for collaboration, while future research directions are discovered. Inferential analysis on independent datasets is a necessary step to further validate results that are obtained.

RQ2 — What technical infrastructure can support reusing Electronic Health Record data for analysis?

A technical infrastructure, consisting of appropriate hardware and software components, can help support reusing EHR data for analysis. Such an infrastructure should address technical, organizational, and ethical challenges that are associated with secondary analysis of EHRs. Based on interviews with relevant experts within the University Medical Center Utrecht (UMCU), including data experts and domain experts, we identified nine requirements:

1. Integrate data sources
2. Preprocess data
3. Store data
4. Support various software and tooling packages
5. Support collaboration and documentation
6. Enhance repeatability
7. Enhance privacy and security
8. Automate data process
9. Support analysis applications

Using the commonly employed Data Warehouse (DWH) model of Inmon (2002) as a starting point, we iteratively combined, refined, split, and removed layers

to finally construct our Capable Reuse of EHR Data (CARED) framework. The framework consists of five layers that process data: an extract layer, a privacy layer, a general preparation layer, a specific preparation layer, and an application layer. Splitting the data preparation over two layers with their own data repositories allows optimal balance between flexibility and time-efficiency for individual researchers. A sixth control layer, consisting of a code base, a scheduling and logging component, and operating system containerization, governs the data process. The framework is designed for researchers to optimally benefit from each others' work and knowledge, while maintaining flexibility in performing analysis.

We finally implemented an infrastructure based on the CARED framework in the Department of Psychiatry of the University Medical Center Utrecht (UMCU) during a period of 6 months, and another 6 months were used to fill the codebase. Obtaining reliable measures of performance before and after implementation proved difficult, because of many confounding effects due to changes that happened simultaneously. However, implementation in the department demonstrates its feasibility, further indicated by various analysis cases that it currently enables.

Conclusion II — A technical infrastructure for reusing EHR data for analysis that is designed along the CARED framework satisfies critical requirements, regarding technical, organizational, and ethical challenges of analyzing EHRs. Such an infrastructure is able to support various analysis applications within a health care organization. Using open source software, such as Python, R, Docker, and GitLab, has the potential to both decrease cost of implementation and increase user adoption.

RQ3 — To what extent can clinical text written in Dutch automatically be de-identified?

Clinical text may hold valuable information, but must be de-identified before it can be used for research purposes, both for legal reasons, and to protect patient privacy. We designed and implemented a pattern-matching based De-identification Method for Dutch Medical Text (DEDUCE), to study whether a pattern-matching based approach can successfully de-identify clinical text written in Dutch. A de-identification method for the Dutch language was not yet available. We obtained a dataset consisting of nurse notes and treatment plans, and split it into a training and testing corpus of 2,000 and 400 examples respectively. Regarding what information to annotate and remove, we found

that 18 Protected Health Identifiers (PHIs) defined in the HIPPAA legislation of the USA are applicable in our case as well. We then used a survey among practitioners to select a subset of relevant PHIs present in our dataset:

1. Person names, including initials
2. Geographical locations smaller than a country
3. Names of institutions that are related to patient treatment
4. Dates
5. Ages
6. Patient numbers
7. Telephone numbers
8. E-mail addresses and URLs

We opted to use a pattern-matching based de-identification method, which has two main advantages over a machine learning approach. Firstly, there is no need for a large annotated corpus, which is not available at present, and secondly, pattern-matching based methods have been shown to have greater generalizability and customizability. DEDUCE, based on rules, lookup lists, and basic pattern matching logic, achieves a macro-averaged F_1-score of 0.862, a recall for person names of 0.964, and missed no person names when evaluated on the test corpus. This result is in line with results in other languages. Although we were not able to test how well DEDUCE generalizes to clinical text outside our corpus, it strives to be applicable to Dutch clinical text in general. A Python implementation of DEDUCE is made available for researchers that work with clinical text.

Conclusion III — Clinical text that is written in Dutch can successfully be de-identified using a method based on rules, lookup-lists, and basic pattern matching logic. Even with perfect accuracy, a de-identification method should in the first place be regarded as a tool to mitigate risk of re-identification.

RQ4 — What Machine Learning techniques are useful for predicting inpatient violence?

To predict inpatient violence, a significant problem in many psychiatric hospitals, we found that clinical text holds most valuable information at the start of admission. We then used an experimental design to find the most useful machine learning techniques to predict inpatient violence based on clinical text. A dataset consisting of 2,521 admissions, and the clinical text available at the start of admission was used. By surveying literature, the most commonly used techniques for text representation are based on bag-of-words, and more recently based on representations learned in an unsupervised way. For text classification, classical models such as Naive Bayes and Support Vector Machines are often used, and novel deep learning models such as Convolutional Neural Networks and Recurrent Neural Networks (RNNs) show high promise. The experimental design we used included a grid search for optimizing hyperparameters, and a cross-validation setup for estimating performance. This setup is useful for comparing different algorithms, but also potentially introduces some bias in the final reported performance.

Regarding representation of text, there is little difference in the bag-of-words weighting scheme, and in the use of text embeddings at the word or document level. However, using text embeddings generally results in higher performance. Regarding the different models, of the two Deep Learning models especially the RNN shows excellent performance, however, it is closely followed by the classical Support Vector Machine algorithm. Training a deep learning model in combination with text embeddings at the word level proved infeasible, probably due to undertraining. Results also show that prediction of violence incidents using machine learning techniques is promising, but the experimental setup used does not provide an objective estimate of predictive validity.

Conclusion IV — Deep learning techniques, both to represent and to classify text, give a small but consistent performance improvement over classical machine learning techniques when predicting violence incidents, even with a modestly sized dataset of clinical text. Approaching this problem using machine learning is promising, but a more rigorous evaluation is needed to prove its use in practice.

RQ5 — How can automatic inpatient violence risk assessment using textual data contribute to the psychiatric practice?

After Chapter 5 established that machine learning is a promising approach in predicting inpatient violence, we used two independent datasets to establish the predictive validity and generalizability of trained models. The two

datasets, again consisting of admissions, violence incidents, and clinical notes, were similar in size and nature, but collected from two different psychiatric health care providers. Based on exploratory analysis at the word level, we found terms that are associated with violence in the first four weeks of admission, either directly related to violence (e.g. *aggressive*, *angry*), and concerning behavioral cues potentially indirectly related to violence (e.g. *reacts*, *walks*). Although many associations were statistically significant after correcting for the family wise error rate, all correlation coefficients were weak, showing that performing violence risk assessment solely based on such keywords is most likely not feasible.

We then used the paragraph2vec algorithm in combination with the Support Vector Machine algorithm to train machine learning models that are able to assess violence risk, in both datasets independently. We measured the predictive validity of models internally using a nested cross validation setup, and exchanged trained models to determine how well their performance generalizes to the other dataset. We found high internal predictive validity, AUC = 0.80 and 0.76 respectively, with no significant difference existing between the two sites. The predictive validity of exchanged models however, AUC = 0.72 and 0.64 respectively, was significantly lower in both cases. This shows us that violence risk assessment using machine learning is possible, with performance that matches or outperforms that of current assessment methods. This finding can potentially be employed in the everyday practice of psychiatric care, because it relies on data that is already routinely captured, an thus requires no additional administrative load. However, using a trained model in another context decreases its predictive validity, losing its advantage over current methods.

Conclusion V — Inpatient violence risk assessment based on applying machine learning techniques to clinical text is possible with good predictive validity, in the same range as existing assessment methods. Therefore, it has high potential for supporting psychiatric care in practice. However, trained models show significantly lower performance when tested on independent datasets. Automatic violence risk assessment based on clinical text might be successful due to its ability to account for behavioral cues and social interaction more easily, compared to existing methods.

RQ6 — How can robust subgroups of psychiatric patients be identified based on Electronic Health Record data?

Definitions of patient subgroups are useful to support clinical care. However,

the current standard for diagnosis based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) lacks validity, and data-driven approaches using various clustering algorithms have not yet agreed on the number and nature of psychiatric patient subgroups. We proposed to use cluster ensembles, combinations of multiple clustering algorithms, to mitigate this problem. Since cluster ensembles can be regarded as more complex than single clustering algorithms, and are less often described in literature, we described a Meta Algorithmic Model (MAM) to support individual researchers in using cluster ensembles. The MAM consists of the Modeling and Evaluation phases of CRISP-DM, applied to cluster ensembles, and describes all process steps that need to be taken, along with the concepts that need to be instantiated in each process step.

Secondly, we applied the cluster ensemble MAM to a dataset of 1,098 Youth Self Report (YSR) questionnaires, answered by adolescents that were treated at the Department of Psychiatry of the UMCU, in order to identify interesting subgroups. We found that individual clustering algorithms indeed obtain different partitions of data, and therefore used three algorithms that obtained diverse partitions. The cluster ensemble approach found three significant clusters. A validation based on inspection of YSR outcome scales and integration with clinical notes showed that these clusters can best be described as *depressive symptoms*, *behavioral problems*, and *low severity*. We then integrated the DSM diagnosis, and found that diagnoses are typically present in multiple clusters, although usually most prominent in a single cluster. Integration with other clinically relevant variables finally showed significant differences in length of treatment and Global Assessment of Functioning at start and end of treatment, showing potential to inform treatment.

Conclusion VI — Finding psychiatric patient subgroups using different clustering algorithms obtains different patient stratifications, in line with existing research. Cluster ensembles are one method to combat high variance in the number and nature of reported patient subgroups, and thus increase robustness of discovered stratifications. Based on application of cluster ensembles to a dataset of Youth Self Reports, three significant subgroups of adolescent patients exist, which differ significantly in several relevant aspects.

8.2 Research Validity

In this section, we will discuss the validity and limitations of the research in this dissertation. First, we will restate some of the threats to validity of analyzing EHR data in general, and then we will discuss the validity of our research approach and research methods used.

8.2.1 Threats to Validity of Analyzing EHR Data

Although using EHR data for analysis has great potential to improve care, it also has several limitations. Most of them have been described or at least touched upon in the previous chapters, but for completeness, four key threats to validity are summarized here.

Data accuracy — given the time-constrained, complex, and sometimes chaotic environment in which EHR data are registered, there is inevitably some noise in these data. The question arises whether EHR data can always be regarded as accurate. On the one hand, EHRs are used to deliver care to patients, and caregivers thus have an incentive to accurately register information. On the other hand, in contrast to traditional research datasets, there are no research protocols that govern the accuracy of EHR data, and any errors that were left uncorrected by the end of treatment are likely to remain unnoticed. There are currently no existing methods that assess the accuracy of datasets composed from EHR data, other than manual curation—which compromises some of the large benefits of using such datasets in the first place. In this work, we have worked against the accuracy issue by always using reasonably sized datasets, in hope of eliminating noise in individual data points through sample size. Still, it is possible for inaccuracies to have an effect on reported outcomes, warranting some caution when interpreting our results.

Data processing — aside from accuracy of registered data, EHR data is subject to many preprocessing steps before analysis can take place. These preprocessing steps are often performed by researchers—rather than practitioners—with a limited understanding of the origins and nature of the data. It is easy to make assumptions about data in selecting, transforming, and processing it, but ascertaining these assumptions are correct is difficult. Thus, additional inaccuracies can be introduced in the steps between registration and analysis. In this work, we have tried to mitigate this issue by manually checking random samples of datasets in collaboration with practitioners, to make a comparison with the EHR they are sourced from, and to identify any other discrepancies. The CARED framework introduced in Chapter 3 is one attempt to further minimize the impact of this problem.

Availability bias — One of the advantages of using EHR data for research is that it is not limited by participant enrollment. This advantage has a downside: the data available in EHRs is clearly biased towards unhealthy people. Where traditional studies have the option to include a control group of healthy subjects in their design, data on healthy subjects is typically not available in EHRs, simply because they do not undergo any treatment. Moreover, patients might continue treatment in other health care organizations, not necessarily due to the nature of disease. This limits the ability of knowledge discovery using EHRs, for example to find new information about the aetiology and early diagnosis of disease, and about patients' disease trajectories. In the end, this entails that the clinical population, rather than the entire population, should be kept in mind when interpreting conclusions that are reached.

Selection bias — even within EHR datasets, making selections is sometimes inevitable. For example, a certain fraction of patients might have missing data regarding a relevant variable, leading to potential exclusion of a non-negligible subset of patients from the dataset. It is not always reasonable to assume that data is missing at random. For example, the group of patients without a diagnosis after the first week of admission might contain a higher proportion of patients with complex pathology, because diagnosis for these patients cannot be established within a week. If there is a specific reason why this data is missing for a specific part of the patient population, excluding these patients can bias the results in some direction. We have tried to mitigate this issue by avoiding making strict patient selections, for example by imputing missing values whenever possible.

8.2.2 Threats to Validity of this Research

Regarding the research approach and research methods that were used, we discuss two key threats to the validity of this dissertation.

External validity — much of the research in this dissertation is carried out within the Department of Psychiatry of the UMCU. This introduces an obvious limitation to the external validity of our research. Aside from using a second independent dataset in Chapter 6, we have not performed external validations of our research artefacts and knowledge discovery outcomes. There might be several reasons why conclusions do not apply in different medical domains, and in different health care organizations within and outside the Netherlands. In different medical domains, different symptoms, underlying causes, and medical protocols exist, which most likely have implications for the types of data that are gathered and the conclusions that can be drawn from them. Different health care organizations within the Netherlands may treat different types of

clinical populations (i.e. with a different distribution of diagnoses), and have practitioners with different opinions regarding knowledge discovery. Health care organizations outside the Netherlands might even be more distinct, in terms of social and cultural differences. In short, under which conditions our findings may generalize to other settings is largely unknown, and would make an interesting topic for future research.

Internal validity — the research artefacts that were described in this dissertation were often designed and validated based on input of domain experts. Given the fact that research into knowledge discovery in EHRs is still only gaining in traction, we often chose to make use of rather exploratory methods for eliciting and validating these artefacts. Although we have done our best to select a representative group of participants for interviewing, surveying, and participation in focus groups, the number of participants was usually modest, again leaving openings for some information potentially left underreported. More rigorous approaches were not always available to researchers, but might have been able to uncover additional information that might serve for further refinement and improvement of the artefacts proposed.

8.3 Future Research

This dissertation has contributed to overcoming some of the barriers of performing knowledge discovery in EHRs, yet this topic of research is still gaining in traction. Application of knowledge discovery techniques to EHRs to directly improve care, of which we have demonstrated the feasibility, can even more be regarded as in its infancy. Plenty of directions for future research can be imagined, and we will describe four of them below.

Deployment in psychiatric practice — despite our efforts, we did not manage to implement one of the results of the second part of this dissertation on the work floor within the duration of this research. It is not hard to imagine several more challenges that arise when attempting to deploy a result on the work floor, to which further research may point out the appropriate solutions. Such research could for example investigate how discovered knowledge can be presented to caregivers in the EHR, how decision support can be implemented in the EHR, and how algorithmic decisions can be appropriately explained. Even after potential implementation, further research should investigate whether such tools actually provide benefit in practice, for example by researching if automatic violence risk assessment such as described in Chapter 6 actually helps decrease inpatient violence.

Integration with healthy subjects data — one of the limitations of using

EHRs, as described in Section 8.2.1, is the fact that this type of data is biased towards non-healthy subjects. In order to enable prevention of disease or diagnosis in very early stages, data about healthy subjects is required in addition to EHR data. In an ever digitizing society, such data is increasingly collected by healthy subjects themselves, for example in the form of wearable devices, health applications, and social media entries. Integrating such data, both in prospective and retrospective study designs, can potentially teach us many new things about the aetiology of disease. How such data should be incorporated is however not yet known. Further research should point out how privacy issues, such as ownership of data and informed consent, and technical issues, such as integrating and analyzing data with measurements from different methods, should be addressed.

Large scale analytics — at the start of this research, one of our long term ambitions was to integrate data of multiple health care organizations, to increase sample size and external validity of analysis. Issues surrounding patient privacy—rightly—hindered this ambition, which was then further impeded by the introduction of the General Data Regulation and Protection (GDPR) act in 2018. However, in order to really improve our fundamental understanding of disease in places where current research falls short, increasing sample size to fully address these complex issues is inevitable. Further research may help find a way forward to obtain larger datasets, if such a way exists, by finding common ground between legal requirements for health care organizations, privacy concerns of human participants, and sample size needs of researchers. If such a way forward does not exist, further research should be devoted to investigate how collaboration between health care organizations is possible without sharing data. Possible options are for instance the use of transfer learning techniques, and structurally replicating others' findings to increase validity.

Improved NLP — our work has demonstrated the importance of clinical text in the care process, and its potential for knowledge discovery. Current NLP methods are however not always applicable, because of the domain specific clinical text barrier and the Dutch language barrier. Increasing the level of both clinical NLP and Dutch NLP, as well as the intersection of both, will enable answering many more research questions. In the future, better methods should be able to accurately recognize information both explicit and implicit within clinical text. In the first case, information extraction makes information such as a patient's symptoms, demographics, and interactions available in a structured format—to the extent this information is suitable for structuring—so that it can be linked to the various other data in the patient's EHR. In the second case, latent information that is not explicitly mentioned can be derived from clinical text. Violence risk assessment is one example of such an

approach, and many more can become feasible when NLP methods are further improved.

8.4 Personal Reflection

When I started this research four years ago, in retrospect, I did not have much sense of what was coming. Academics work in an environment full of scrutiny and setbacks, all the while striving for the highest quality under an ingrain lack of time. At times, completing this dissertation felt like quite the endeavor, and at most other times it felt like a privilege. In these final paragraphs, I will take the opportunity to reflect on this research and on some of the lessons learned during it.

Domain understanding — with a background in artificial intelligence, at the start of this research I entered two domains that were largely new to me. It took me quite some time to get accustomed to *information science*, the field within which I did my research, where completely different research methods than I was used to are employed. Although I would still not primarily consider myself an *information scientist*, after some adjustment I have definitely come to appreciate the organizational aspects of computing science. The domain of psychiatry was clearly even further away from my own expertise, but becoming accustomed to the organization, terminology, and research topics proved quite feasible. I believe to have acquired some rudimentary sense of what psychiatry is through many interactions, yet, as an outsider, the everyday practice has remained somewhat mysterious. Herein lies the main motivation for proper engagement between physicians and technicians: only through interaction, collaboration, and education will knowledge discovery truly fulfill its potential to improve care.

On the fence — despite how far science has come, in today's world plenty of new questions are discovered every day. An ideal world would allow ample consideration to test the feasibility of many ideas and research directions, but alas, academia has deadlines to be met, classes to be taught, and papers to be published. Especially for young researchers, deciding which idea to pursue next can be a daunting task. This work has stressed exploratory research in EHRs as an important benefit of such datasets. However, when presented with a large dataset of EHRs, coming up with the most promising research directions in a medical domain in which I am no expert formed quite a challenge. Fortunately, I always found myself among many colleagues and practitioners who gladly helped to distinguish the promising and interesting from the unlikely and ordinary. In the end, this research has seen both a fair share of

fruitful and abandoned projects, however, I am glad to note the total number of failures has largely been kept in check.

Per aspera ad astra — as a strong proponent of always keeping societal impact in mind, the medical research domain suits me well. My initial enthusiasm for working in academia at some point faded for some skepticism whether scientific research was really the way to maximize my impact, finally coming round a second time during the last parts of this research. Although the real world pay-off of one's work can sometimes seem obscure, research matters. For this reason, many interactions I had with persons in psychiatry have made a lasting impression on me. Both caregivers and patients are faced with an arduous task, and deserve a lot of appreciation for doing their utmost to improve well-being every day. My hope is that this work has made a positive contribution, however small, to a future in which health care can even better fulfill the needs of caregivers and patients.

Bibliography

Abbe, A., Grouin, C., Zweigenbaum, P., and Falissard, B. (2015). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research*, 25(2):86–100.

Abderhalden, C., Needham, I., Dassen, T., Halfens, R., Haug, H.-J., and Fischer, J. E. (2008). Structured risk assessment and violence in acute psychiatric wards: Randomised controlled trial. *British Journal of Psychiatry*, 193(01):44–50.

Adam, D. (2013). Mental health: On the spectrum. *Nature*, 496(7446):416–418.

Adler-Milstein, J., DesRoches, C. M., Kralovec, P., et al. (2015a). Electronic health record adoption in US hospitals: Progress continues, but challenges persist. *Health Affairs*, 34(12):2174–2180.

Adler-Milstein, J., Everson, J., and Lee, S.-Y. D. (2015b). EHR adoption and hospital performance: Time-related effects. *Health Services Research*, 50(6):1751–1771.

Ægisdóttir, S., White, M. J., Spengler, P. M., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3):341–382.

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Springer US.

Alsaleem, S. (2011). Automated arabic text categorization using SVM and NB. *International Arab Journal of eTechnology*.

American Psychiatric Association (1952). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Mental Hospital Service.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Amore, M., Menchetti, M., Tonti, C., et al. (2008). Predictors of violent behavior among acute psychiatric patients: Clinical study. *Psychiatry and Clinical Neurosciences*, 62(3):247–255.

Andreasen, N. C. (1985). *The Broken Brain: The Biological Revolution in Psychiatry*. Harper Collins, New York, USA.

Angus, D. C. (2015). Fusing randomized trials with big data. *JAMA*, 314(8):767.

Appleby, J. (2012). Rises in healthcare spending: Where will it end? *BMJ*, 345:e7127–e7127.

Apte, M., Neidell, M., Furuya, E. Y., Caplan, D., Glied, S., and Larson, E. (2011). Using electronically available inpatient hospital data for research. *Clinical and Translational Science*, 4(5):338–345.

Azevedo, A. I. R. L. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In *IADIS European Conference Data Mining*.

Badawi, O., Brennan, T., Celi, L. A., et al. (2014). Making big data useful for health care: A summary of the inaugural MIT critical data conference. *JMIR Medical Informatics*, 2(2):e22.

Banerjee, A. and Dave, R. (2004). Validating clusters using the Hopkins statistic. In *2004 IEEE International Conference on Fuzzy Systems*. IEEE.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and 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.

Bauer, C. R. K. D., Ganslandt, T., Baum, B., et al. (2016). Integrated data repository toolkit (IDRT). *Methods of Information in Medicine*, 55(02):125–135.

Beckett, J. (2017). Evaluating some of the approaches: Biomedical versus alternative perspectives in understanding mental health. *Journal of Psychiatry and Psychiatric Disorders*, 1(2):103–107.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Lecture Notes in Computer Science*, pages 437–478. Springer Berlin Heidelberg.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Beynon-Davies, P., Tudhope, D., and Mackay, H. (1999). Information systems prototyping in practice. *Journal of Information Technology*, 14(1):107–120.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Bogner, A., Littig, B., and Menz, W. (2009). Introduction: Expert interviews — an introduction to a new methodological debate. In *Interviewing Experts*, pages 1–13. Palgrave Macmillan UK, London.

Boongoen, T. and Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28:1–25.

Borking, J. J. and Raab, C. D. (2001). Laws, PETs and other technologies for privacy protection. *Journal of Information, Law and Technology*, 1:1–14.

Borum, R., Bartel, P. A., and Forth, A. E. (2005). *Mental health screening and assessment in juvenile justice*, chapter Structured Assessment of Violence Risk in Youth, pages 311–323. The Guilford Press.

Bram, J. T., Warwick-Clark, B., Obeysekare, E., and Mehta, K. (2015). Utilization and monetization of healthcare data in developing countries. *Big Data*, 3(2):59–66.

Brennan, P. F. and Bakken, S. (2015). Nursing needs big data and big data needs nursing. *Journal of Nursing Scholarship*, 47(5):477–484.

Britten, N. (1995). Qualitative interviews in medical research. *BMJ*, 311(6999):251–253.

Brown, T., Di Nardo, P., Lehman, C., and Campbell, L. (2001). Reliability of DSM-IV anxiety and mood disorders: Implications for the classification of emotional disorders. *Journal of Abnormal Psychology*, 110(1):49–58.

Campbell, M. A., French, S., and Gendreau, P. (2009). The prediction of violence in adult offenders. *Criminal Justice and Behavior*, 36(6):567–590.

Chapman, P., Clinton, J., Kerber, R., et al. (2000). CRISP-DM 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Chassin, M. R. and Galvin, R. W. (1998). The urgent need to improve health care quality. Institute of Medicine national roundtable on health care quality. *JAMA*, 280(11):1000–5.

Chawla, N. V. and Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(S3):660–665.

Chen, Chiang, and Storey (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165.

Chen, C.-Y., Lee, P. H., Castro, V. M., et al. (2018). Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. *Translational Psychiatry*, 8(1).

Cheruvelil, K. S., Soranno, P. A., Weathers, K. C., et al. (2014). Creating and maintaining high-performing collaborative research teams: The importance of diversity and interpersonal skills. *Frontiers in Ecology and the Environment*, 12(1):31–38.

Chiu, B., Crichton, G., Korhonen, A., and Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174. Association for Computational Linguistics.

Cohen, J. J. and Siegel, E. K. (2005). Academic medical centers and medical research. *JAMA*, 294(11):1367.

Cole, V. T., Apud, J. A., Weinberger, D. R., and Dickinson, D. (2012). Using latent class growth analysis to form trajectories of premorbid adjustment in schizophrenia. *Journal of Abnormal Psychology*, 121(2):388–395.

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, 350:g7594–g7594.

Collins, P. Y., Patel, V., Joestl, S. S., et al. (2011). Grand challenges in global mental health. *Nature*, 475(7354):27–30.

Coorevits, P., Sundgren, M., Klein, G. O., et al. (2013). Electronic health records: New opportunities for clinical research. *Journal of Internal Medicine*, 274(6):547–560.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cristea, I. A., Karyotaki, E., Hollon, S. D., Cuijpers, P., and Gentili, C. (2019). Biological markers evaluated in randomized trials of psychological treatments for depression: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 101:32–44.

Cuthbert, B. N. and Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, 11(1).

Dack, C., Ross, J., Papadopoulos, C., Stewart, D., and Bowers, L. (2013). A review and meta-analysis of the patient factors associated with psychiatric in-patient aggression. *Acta Psychiatrica Scandinavica*, 127(4):255–268.

Dalal, M. K. and Zaveri, M. A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications*, 28(2):37–40.

Danciu, I., Cowan, J. D., Basford, M., et al. (2014). Secondary use of clinical data: The Vanderbilt approach. *Journal of Biomedical Informatics*, 52:28–35.

Deacon, B. J. (2013). The biomedical model of mental disorder: A critical analysis of its validity, utility, and effects on psychotherapy research. *Clinical Psychology Review*, 33(7):846–861.

Deacon, B. J. and Lickel, J. J. (2009). On the brain disease model of mental disorders. *The Behavior Therapist*, 32(6):113–118.

Dean, B. B., Lam, J., Natoli, J. L., Butler, Q., Aguilar, D., and Nordyke, R. J. (2009). Review: Use of electronic medical records for health outcomes research. *Medical Care Research and Review*, 66(6):611–638.

Dean, C. E. (2017). Social inequality, scientific inequality, and the future of mental illness. *Philosophy, Ethics, and Humanities in Medicine*, 12(1):1–10.

Deleger, L., Molnar, K., Savova, G., et al. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845.

Deshpande, V. P., Erbacher, R. F., and Harris, C. (2007). An evaluation of Naive Bayesian anti-spam filtering techniques. In *2007 IEEE SMC Information Assurance and Security Workshop*. IEEE.

DesRoches, C. M., Charles, D., Furukawa, M. F., et al. (2013). Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. *Health Affairs*, 32(8):1478–1485.

DiCicco-Bloom, B. and Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, 40(4):314–321.

Dollfus, S., Everitt, B., Ribeyre, J. M., Assouly-Besse, F., Sharp, C., and Petit, M. (1996). Identifying subtypes of schizophrenia by cluster analyses. *Schizophrenia Bulletin*, 22(3):545–555.

Double, D. (2002). The limits of psychiatry. *BMJ*, 324(7342):900–904.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., and Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, version 3 (HCR-20v3): Development and overview. *International Journal of Forensic Mental Health*, 13(2):93–108.

Douglass, M., Cliffford, G., Reisner, A., Long, W., Moody, G., and Mark, R. (2005). De-identification algorithm for free-text nursing notes. *Computers in Cardiology*, pages 331–334.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., and Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical Teacher*, 36(3):216–222.

Elo, S. and Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115.

Emam, K. E., Jabbouri, S., Sams, S., Drouet, Y., and Power, M. (2006). Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, 8(4):e28.

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

European Data Protection Directive (1995). Directive 95/46/ECV of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. the MIT Press.

Fazel, S., Singh, J. P., Doll, H., and Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ*, 345:e4692–e4692.

Fenz, S., Heurix, J., Neubauer, T., and Rella, A. (2014). De-identification of unstructured paper-based health records for privacy-preserving secondary use. *Journal of Medical Engineering & Technology*, 38(5):260–268.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, 15(1).

Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., and Meystre, S. M. (2012a). BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association*, 20(1):77–83.

Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., and Meystre, S. M. (2012b). Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. *BMC Medical Research Methodology*, 12(1).

Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., and Brown, J. S. (2014). Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4):578–582.

Ford, E., Stockdale, J., Jackson, R., and Cassell, J. (2017). For the greater good? patient and public attitudes to use of medical free text data in research. *International Journal of Population Data Science*, 1(1).

Forman, G. and Scholz, M. (2010). Apples-to-apples in cross-validation studies. *ACM SIGKDD Explorations Newsletter*, 12(1):49.

Fountain, C., Winter, A. S., and Bearman, P. S. (2012). Six developmental trajectories characterize children with autism. *PEDIATRICS*, 129(5):e1112–e1120.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, 13(3):57–70.

Fred, A. (2001). Finding consistent clusters in data partitions. In *Proceedings of the 3rd International Workshop on Multiple Classifier Systems*, pages 309–318. Springer.

Friedlin, F. J. and McDonald, C. J. (2008). A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

Friedman, C., Rubin, J., Brown, J., et al. (2015). Toward a science of learning systems: A research agenda for the high-functioning Learning Health System. *Journal of the American Medical Informatics Association*, 22(1):43–50.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. Technical report, Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

Garla, V., Taylor, C., and Brandt, C. (2013). Semi-supervised clinical text classification with laplacian SVMs: An application to cancer case management. *Journal of Biomedical Informatics*, 46(5):869–875.

George, J., KUmar, B. V., and Kumar, V. S. (2015). Data warehouse design considerations for a healthcare business intelligence system. In *Proceedings of the World Congress on Engineering 2015, London, U.K.*

Ghaemi, R., Sulaiman, N., Ibrahim, M., and Mustapha, N. (2009). A survey: Clustering ensembles techniques. *International Journal of Computer and Information Engineering*, 3(2):365–374.

Gil, Y., Cheung, W. K., Ratnakar, V., and Chan, K.-k. (2007). Privacy enforcement in data analysis workflows. In *Proceedings of the 2007 International Conference on Privacy Enforcement and Accountability with Semantics - Volume 320, Busan, Korea*, pages 41–48.

Gill, P., Stewart, K., Treasure, E., and Chadwick, B. (2008). Methods of data collection in qualitative research: Interviews and focus groups. *British Dental Journal*, 204(6):291–295.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

Goodman, A., Pepe, A., Blocker, A. W., et al. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4):e1003542.

Grouin, C., Rosier, A., Dameron, O., and Zweigenbaum, P. (2009). Testing tactics to localize de-identification. *Studies in health technology and informatics*, 150:735–739.

Groves, P., Kayyali, B., Knott, D., and van Kuiken, S. (2013). The 'big data' revolution in healthcare: Accelerating value and innovation. McKinsey Global Institute.

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402.

Hagoort, K., Menger, V., Velders, F., Deschamps, P., and Scheepers, F. E. (2018). 29.3 application of new data sources: Text mining and story banking. *Journal of the American Academy of Child & Asolescent Psychiatry*, 57(10):S314.

Hajian-Tilaki, K. O. and Hanley, J. A. (2002). Comparison of three methods for estimating the standard error of the area under the curve in ROCAnalysis of quantitative data. *Academic Radiology*, 9:1278–1285.

Halligan, S., Altman, D. G., and Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology*, 25(4):932–939.

Hammami, R., Bellaaj, H., and Kacem, A. H. (2014). Interoperability of healthcare information systems. In *The 2014 International Symposium on Networks, Computers and Communications*. IEEE.

Hammerla, N. Y., Halloran, S., and Plötz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. In

Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, pages 1533–1540. AAAI Press.

Harpaz, R., Callahan, A., Tamang, S., et al. (2014). Text mining for adverse drug events: The promise, challenges, and state of the art. *Drug Safety*, 37(10):777–790.

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6):836–850.

Hazlehurst, B. L., Kurtz, S. E., Masica, A., et al. (2015). CER Hub: An informatics platform for conducting comparative effectiveness research using multi-institutional, heterogeneous, electronic clinical data. *International Journal of Medical Informatics*, 84(10):763–773.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hersh, W., Cimino, J., Payne, P. R., et al. (2013a). Recommendations for the use of operational electronic health record data in comparative effectiveness research. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(1):14.

Hersh, W. R., Weiner, M. G., Embi, P. J., et al. (2013b). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*, 51:S30–S37.

Higgins, N., Watts, D., Bindman, J., Slade, M., and Thornicroft, G. (2005). Assessing violence risk in general adult psychiatry. *Psychiatric Bulletin*, 29(04):131–133.

Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., et al. (2000). Reliability and validity of DSM-IV Axis V. *American Journal of Psychiatry*, 157(11):1858–1863.

HIPAA (1996). Health insurance portability and accountability act.

HITECH (2009). Health information technology for economic and clinical health act.

Hripcsak, G. and Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121.

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., and Melton, G. (2011). Bias associated with mining electronic health records. *Journal of Biomedical Discovery and Collaboration*, 6:48–52.

Huang, H., Liu, Y., Yuan, M., and Marron, J. S. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993.

Hughes, M., Li, I., Kotoulas, S., and Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Studies in Health Technology and Informatics*, 235(Informatics for Health: Connected Citizen-Led Wellness and Population Health):246–250.

Inmon, W. H. (2002). *Building the Data Warehouse, 3rd Edition.* John Wiley & Sons, Inc., New York, NY, USA.

Inoue, M., Tsukano, K., Muraoka, M., Kaneko, F., and Okamura, H. (2006). Psychological impact of verbal abuse and violence by patients on nurses working in psychiatric departments. *Psychiatry and Clinical Neurosciences*, 60(1):29–36.

Insel, T. R. (2017). Digital phenotyping. *JAMA*, 318(13):1215.

Institute of Medicine (2010). *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary.* National Academies Press, Washington (DC).

Iozzino, L., Ferrari, C., Large, M., Nielssen, O., and de Girolamo, G. (2015). Prevalence and risk factors of violence by psychiatric acute inpatients: A systematic review and meta-analysis. *PLOS ONE*, 10(6):e0128536.

Jacobson, O. and Dalianis, H. (2016). Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. In *ACL Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP 2016*, pages 191–195.

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

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142. Springer Berlin Heidelberg.

Johnson, K. E., Kamineni, A., Fuller, S., Olmstead, D., and Wernli, K. J. (2014). How the provenance of electronic health record data matters for research: A case example using system mapping. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 2(1):1058.

Kasthurirathne, S. N., Mamlin, B., Kumara, H., Grieve, G., and Biondich, P. (2015). Enabling better interoperability for healthcare: Lessons in developing a standards based application programing interface for electronic medical record systems. *Journal of Medical Systems*, 39(11).

Katal, A., Wazid, M., and Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)*. IEEE.

Kathol, R. G., Kunkel, E. J., Weiner, J. S., et al. (2009). Psychiatrists for medically complex patients: Bringing value at the physical health and mental health/substance-use disorder interface. *Psychosomatics*, 50(2):93–107.

Kattan, M. W. (2011). Factors affecting the accuracy of prediction models limit the comparison of rival prediction models when applied to separate data sets. *European Urology*, 59(4):566–567.

Khoury, M. J. and Ioannidis, J. P. A. (2014). Big data meets public health. *Science*, 346(6213):1054–1055.

King, J., Patel, V., Jamoom, E. W., and Furukawa, M. F. (2013). Clinical benefits of electronic health record use: National findings. *Health Services Research*, 49(1pt2):392–404.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*.

Kitzinger, J. (1995). Qualitative research. introducing focus groups. *BMJ*, 311(7000):299–302.

Kleinman, A., Caetano, S. C., Brentani, H., et al. (2014). Attention-based classification pattern, a research domain criteria framework, in youths with bipolar disorder and attention-deficit/hyperactivity disorder. *Australian & New Zealand Journal of Psychiatry*, 49(3):255–265.

Koh, H. C., Tan, G., et al. (2005). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65.

Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428.

Korde, V. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85–99.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Kraemer, H. C. and Freedman, R. (2014). Computer aids for the diagnosis of anxiety and depression. *American Journal of Psychiatry*, 171(2):134–136.

Krishna, R., Kelleher, K., and Stahlberg, E. (2007). Patient confidentiality in the research use of clinical medical databases. *American Journal of Public Health*, 97(4):654–658.

Krishnankutty, B., Kumar, B. N., Moodahadu, L., and Bellary, S. (2012). Data management in clinical research: An overview. *Indian Journal of Pharmacology*, 44(2):168.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7):1163–1170.

Kuncheva, L. and Hadjitodorov, S. (2004). Using diversity in cluster ensembles. In *2004 IEEE International Conference on Systems, Man and Cybernetics*. IEEE.

Kuper, A. and D'Eon, M. (2010). Rethinking the basis of medical knowledge. *Medical Education*, 45(1):36–43.

Kupwade Patil, H. and Seshadri, R. (2014). Big data security and privacy issues in healthcare. In *2014 IEEE International Congress on Big Data*. IEEE.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lahti, A. C. (2016). Making progress toward individualized medicine in the treatment of psychosis. *American Journal of Psychiatry*, 173(1):5–7.

Lakhan, S. E., Vieira, K., and Hamlat, E. (2010). Biomarkers in psychiatry: Drawbacks and potential for misuse. *International Archives of Medicine*, 3(1):1.

Lan, M., Tan, C.-L., Low, H.-B., and Sung, S.-Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Special interest tracks and posters of the 14th international conference on World Wide Web - WWW'05*. ACM Press.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, pages II–1188–II–1196. JMLR.org.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, C. H. and Yoon, H.-J. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice*, 36(1):3–11.

Lee, E. S., Black, R. A., Harrington, R. D., Tarczy-Hornoch, P., and FACMI (2015). Characterizing secondary use of clinical data. In *Proceedings of the AMIA Joint Summits on Translational Science 2015*.

Lee, J., McCullough, J. S., and Town, R. J. (2013). The impact of health information technology on hospital productivity. *The RAND Journal of Economics*, 44(3):545–568.

Lee, W., Bindman, J., Ford, T., et al. (2007). Bias in psychiatric case–control studies: literature survey. *British Journal of Psychiatry*, 190(03):204–209.

Lewandowski, K. E., Sperry, S. H., Cohen, B. M., and Öngür, D. (2014). Cognitive variability in psychotic disorders: A cross-diagnostic cluster analysis. *Psychological Medicine*, 44(15):3239–3248.

Li, H., Li, X., Ramanathan, M., and Zhang, A. (2014). Identifying informative risk factors and predicting bone disease progression via deep belief networks. *Methods*, 69(3):257–265.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437.

Liang, Z., Zhang, G., Huang, J. X., and Hu, Q. V. (2014). Deep learning for healthcare decision making with EMRs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Lin, J. and Haug, P. (2006). Data preparation framework for preprocessing clinical data in data mining. In *Proceedings of the Annual AMIA Symposium*.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. C. (2015). Learning to diagnose with LSTM recurrent neural networks. *CoRR*, abs/1511.03677.

Liu, V., Musen, M. A., and Chou, T. (2015). Data breaches of protected health information in the united states. *JAMA*, 313(14):1471.

Liu, Z., Yang, M., Wang, X., et al. (2017). Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, 17(S2).

Lokhandwala, S. and Rush, B. (2016). Objectives of the secondary analysis of electronic health record data. In *Secondary Analysis of Electronic Health Records*, pages 3–7. Springer International Publishing.

Lu, Z. and Su (2010). Clinical data management: Current status, challenges, and future directions from industry perspectives. *Open Access Journal of Clinical Trials*, page 93.

Lv, X., Guan, Y., Yang, J., and Wu, J. (2016). Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248.

Maden, A. (2003). Standardised risk assessment: Why all the fuss? *Psychiatric Bulletin*, 27(06):201–204.

Maenner, M. J., Yeargin-Allsopp, M., Braun, K. V. N., Christensen, D. L., and Schieve, L. A. (2016). Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PLOS ONE*, 11(12):e0168224.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):433–447.

McCarty, C. A., Chisholm, R. L., Chute, C. G., et al. (2011). The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics*, 4(1).

McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., and Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable- and person-focused analyses. *Law and Human Behavior*, 32(4):325–338.

McIntosh, A. M., Stewart, R., John, A., et al. (2016). Data science for mental health: a UK perspective on a global challenge. *The Lancet Psychiatry*, 3(10):993–998.

Meertens Instituut (2016). Netwerk Naamkunde. `http://www.naamkunde.net/?page_id=289`. Accessed December 29, 2016.

Menger, V., Scheepers, F., and Spruit, M. (2018a). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6):981.

Menger, V., Scheepers, F., van Wijk, L. M., and Spruit, M. (2018b). DEDUCE: A pattern matching method for automatic de-identification of dutch medical text. *Telematics and Informatics*, 35(4):727–736.

Menger, V., Spruit, M., de Bruin, J., Kelder, T., and Scheepers, F. (2019a). Supporting reuse of EHR data in healthcare organizations: The CARED research infrastructure framework. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, volume 5: HEALTHINF, pages 41–50. SCITEPRESS - Science and Technology Publications.

Menger, V., Spruit, M., Hagoort, K., and Scheepers, F. (2016). Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Computational and Mathematical Methods in Medicine*, 2016:1–11.

Menger, V., Spruit, M., van der Klift, W., and Scheepers, F. (2019b). Using cluster ensembles to identify psychiatric patient subgroups. In *Artificial Intelligence in Medicine*, pages 252–262. Springer International Publishing.

Menger, V., Spruit, M., van Est, R., Nap, E., and Scheepers, F. (2019c). Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Network Open*, 2(7):e196709.

Meulendijk, M., Spruit, M., van Maanen, C. D., Numans, M., Brinkkemper, S., and Jansen, P. (2013). General practitioners' attitudes towards decision-supported prescribing: An analysis of the Dutch primary care sector. *Health Informatics Journal*, 19(4):247–263.

Meystre, S. M., Ferrández, Ó., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2014). Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 50:142–150.

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., and Lehmann, C. U. (2017). Clinical data reuse or secondary use: Current status and potential future progress. *Yearbook of Medical Informatics*, 26(01):38–52.

Michel-Verkerke, M. B. and Spil, T. A. M. (2002). Electronic patient records in the Netherlands, Luctor et Emergo; but who is struggling and what will emerge? In *Proceedings of the 10th European Conference on Information Systems, Gdansk, Poland*, pages 1443–1453.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Milovic, B. (2012). Prediction and decision making in health care using data mining. *International Journal of Public Health Science (IJPHS)*, 1(2).

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6(1).

Möller, H.-J. (2009). Standardised rating scales in psychiatry: Methodological basis, their possibilities and limitations and descriptions of important rating scales. *The World Journal of Biological Psychiatry*, 10(1):6–26.

Monahan, J. and Skeem, J. L. (2014). The evolution of violence risk assessment. *CNS Spectrums*, 19(05):419–424.

Moor, G. D., Sundgren, M., Kalra, D., et al. (2015). Using electronic health records for clinical research: The case of the EHR4CR project. *Journal of Biomedical Informatics*, 53:162–173.

Morgan, D. (1997). *Focus Groups as Qualitative Research*. SAGE Publications, Inc., London, UK.

Morgan, D. (1998). *The Focus Group Guidebook*. SAGE Publications, Inc., London, UK.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA*, 309(13):1351.

Murphy, S. N., Dubey, A., Embi, P. J., et al. (2012). Current state of information technologies for the clinical research enterprise across academic medical centers. *Clinical and Translational Science*, 5(3):281–284.

Murphy, S. N., Weber, G., Mendis, M., et al. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130.

Nair, S., Hsu, D., and Celi, L. A. (2016). Challenges and opportunities in secondary analyses of electronic health record data. In *Secondary Analysis of Electronic Health Records*, pages 17–26. Springer International Publishing.

Neamatullah, I., Douglass, M. M., wei H Lehman, L., et al. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1).

Nickerson, P., Tighe, P., Shickel, B., and Rashidi, P. (2016). Deep neural network architectures for forecasting analgesic response. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.

Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2–3):103–134.

Nijman, H., Bowers, L., Oud, N., and Jansen, G. (2005). Psychiatric nurses' experiences with inpatient aggression. *Aggressive Behavior*, 31(3):217–227.

Nijman, H. L., Muris, P., Merckelbach, H. L., et al. (1999). The staff observation aggression scale-revised (SOAS-R). *Aggressive Behaviour*, 25:197–209.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13):1216–1219.

Obermeyer, Z. and Lee, T. H. (2017). Lost in thought — The limits of the human mind and the future of medicine. *New England Journal of Medicine*, 377(13):1209–1211.

OCEBM Levels of Evidence Working Group (2011). The Oxford levels of evidence 2. `https://www.cebm.net/index.aspx?o=5653`.

O'Leary, D. E. (1988). Expert system prototyping as a research tool. In *Applied Expert Systems*, pages 17–31. Elsevier Science Publishers, North-Holland, the Netherlands.

Olino, T. M., Klein, D. N., Lewinsohn, P. M., Rohde, P., and Seeley, J. R. (2010). Latent trajectory classes of depressive and anxiety disorders from adolescence to adulthood: Descriptions of classes and associations with risk factors. *Comprehensive Psychiatry*, 51(3):224–235.

Onwubolu, G. (2009). An inductive data mining system framework. In *Proceedings of the International Workshop on Inductive Modeling (IWIM '09)*, pages 108–113.

Oquendo, M. A., Baca-Garcia, E., Artés-Rodríguez, A., et al. (2012). Machine learning and data mining: Strategies for hypothesis generation. *Molecular Psychiatry*, 17(10):956–959.

Ozomaro, U., Wahlestedt, C., and Nemeroff, C. B. (2013). Personalized medicine in psychiatry: Problems and promises. *BMC Medicine*, 11(1).

Pannucci, C. J. and Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery*, 126(2):619–625.

Papadopoulos, C., Ross, J., Stewart, D., Dack, C., James, K., and Bowers, L. (2012). The antecedents of violence and aggression within psychiatric in-patient settings. *Acta Psychiatrica Scandinavica*, 125(6):425–439.

Passos, I. C., Mwangi, B., and Kapczinski, F. (2016). Big data analytics and machine learning: 2015 and beyond. *The Lancet Psychiatry*, 3(1):13–15.

Patel, R., Jayatilleke, N., Broadbent, M., et al. (2015). Negative symptoms in schizophrenia: A study in a large clinical sample of patients using a novel automated method. *BMJ Open*, 5(9):e007619.

Peek, N., Holmes, J. H., and Sun, J. (2014). Technical challenges for big data in biomedicine and health: Data sources, infrastructure, and analytics. *Yearbook of Medical Informatics*, 23(01):42–47.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Pestian, J. P., Brew, C., Matykiewicz, P., et al. (2007). A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP'07*. Association for Computational Linguistics.

Peters, T. E. (2017). Transformational impact of health information technology on the clinical practice of child and adolescent psychiatry. *Child and Adolescent Psychiatric Clinics of North America*, 26(1):55–66.

Pfeffer, C. R., Solomon, G., Plutchik, R., Mizruchi, M. S., and Weiner, A. (1985). Variables that predict assaultiveness in child psychiatric inpatients. *Journal of the American Academy of Child Psychiatry*, 24(6):775–780.

Pollard, T., Dernoncourt, F., Finlayson, S., and Velasquez, A. (2016). Data preparation. In *Secondary Analysis of Electronic Health Records*, pages 101–114. Springer International Publishing.

Priest, E. L., Klekar, C., Cantu, G., et al. (2014). Developing electronic data methods infrastructure to participate in collaborative research networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 2(1):18.

Priyanka, K. and Kulennavar, N. (2014). A survey on big data analytics in health care. *International Journal of Computer Science and Information Technologies*, 5(4):5865–5868.

Probstfield, J. L. and Frye, R. L. (2011). Strategies for recruitment and retention of participants in clinical trials. *JAMA*, 306(16):1798–1799.

Quinsey, V. L., Harris, G. T., Rice, M. E., and Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk.* American Psychological Association.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1).

Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., and Palaniappan, B. (2009). Automatic classification of Tamil documents using vector space model and artificial neural network. *Expert Systems with Applications*, 36(8):10914–10918.

Ramakrishnan, N., Hanauer, D., and Keller, B. (2010). Mining electronic health records. *Computer*, 43(10):77–81.

Rea, S., Pathak, J., Savova, G., et al. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics*, 45(4):763–771.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Reynolds, G. P., McKelvey, J. S., Reinharth, J., et al. (2013). Predictors of persistent aggression on the psychiatric inpatient service. *Comprehensive Psychiatry*, 54(8):e34.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Expaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16*, San Fransisco, California. ACM Press.

Rittenhouse, D. R., Ramsay, P. P., Casalino, L. P., McClellan, S., Kandel, Z. K., and Shortell, S. M. (2017). Increased health information technology adoption and use among small primary care physician practices over time: A national cohort study. *The Annals of Family Medicine*, 15(1):56–62.

Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1).

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., and Johnson, K. B. (2011). Data from clinical notes: A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Roski, J., Bo-Linn, G. W., and Andrews, T. A. (2014). Creating value in health care through big data: Opportunities and policy implications. *Health Affairs*, 33(7):1115–1122.

Ruch, P., Baud, R. H., Rassinoux, A. M., Bouillon, P., and Robert, G. (2000). Medical document anonymization with a semantic lexicon. In *Proceedings of the AMIA Symposium*, pages 729–733.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72.

Safran, C. (2014). Reuse of clinical data. *Yearbook of Medical Informatics*, 23(01):52–54.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

Sarker, A. and Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207.

Savova, G. K., Masanz, J. J., Ogren, P. V., et al. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Scheepers, F., Menger, V., and Hagoort, K. (2018). Datascience in de psychiatrie. *Tijdschrift voor psychiatrie*, 60(3):205–209.

Scheurwegs, E., Luyckx, K., Van der Schueren, F., and Van den Bulcke, T. (2013). De-identification of clinical free text in Dutch with limited training data: A case study. In *Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, Hissar, Bulgaria*, pages 18–23. INCOMA Ltd. Shoumen, BULGARIA.

Shen, J., Lee, P., H., J. H., and Shatkay (2007). Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders. In *AMIA Annual Symposium Proceedings 2007*, volume 2007, pages 666–670.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2018). 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.

Shin, S.-Y., Park, Y. R., Shin, Y., et al. (2015). A de-identification method for bilingual clinical texts of various note types. *Journal of Korean Medical Science*, 30(1):7.

Silverman, W. K., Saavedra, L. M., and Pina, A. A. (2001). Test-retest reliability of anxiety symptoms and diagnoses with the anxiety disorders interview schedule for DSM-IV: Child and parent versions. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40(8):937–944.

Simon, G. E., Unützer, J., Young, B. E., and Pincus, H. A. (2000). Large medical databases, population-based research, and patient confidentiality. *The American journal of psychiatry*, 157:1731–1737. KIE Bib: biomedical research; confidentiality/mental health.

Singh, J. P., , Yang, S., and Mulvey, E. P. (2015). Reporting guidance for violence risk assessment predictive validity studies: The RAGEE statement. *Law and Human Behavior*, 39(1):15–22.

Singh, J. P., Desmarais, S. L., Hurducas, C., et al. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health*, 13(3):193–206.

Singh, J. P., Grann, M., and Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, 31(3):499–513.

Spruit, M. and Jagesar, R. (2016). Power to the people! - Meta-Algorithmic Modelling in applied data science. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and Technology Publications.

Steinert, T. (2002). Prediction of inpatient violence. *Acta Psychiatrica Scandinavica*, 106(s412):133–141.

Strauss, A. and Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publicatoins, Inc., Thousand Oaks, CA, US.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Suchting, R., Green, C. E., Glazier, S. M., and Lane, S. D. (2018). A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Research*, 268:217–222.

Sun, A., Lim, E.-P., and Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, 48(1):191–201.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Clinical intervention prediction and understanding with deep neural networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68, pages 322–337.

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tedre, M. and Moisseinen, N. (2014). Experiments in computing: A survey. *The Scientific World Journal*, 2014:1–11.

Teo, A. R., Holley, S. R., Leary, M., and McNiel, D. E. (2012). The relationship between level of training and accuracy of violence risk assessment. *Psychiatric Services*, 63(11):1089–1094.

Thomas, S. M., Mamlin, B., Schadow, G., and McDonald, C. (2002). A successful technique for removing names in pathology reports using an augmented search and replace method. In *Proceedings of the AMIA Symposium*, pages 777–781.

Tijsen, R., Spruit, M., van de Ridder, M., and van Raaij, B. (2011). BI-FIT: aligning business intelligence end-users, tasks and technologies. In *Enterprise Information Systems Design, Implementation and Management: Organizational Applications*, pages 162–177. IGI Global.

Topchy, A., Jain, A. K., and Punch, W. (2003). Combining multiple weak clusterings. In *Proceedings of the Third IEEE International Conference on Data Mining*, page 331, Washington, DC, USA. IEEE Computer Society.

Topchy, A. P., Law, M. H. C., Jain, A. K., and Fred, A. L. (2004). Analysis of consensus partition in cluster ensemble. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 225–232, Washington, DC, USA. IEEE Computer Society.

Torous, J. and Baker, J. T. (2016). Why psychiatry needs data science and data science needs psychiatry. *JAMA Psychiatry*, 73(1):3.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7):1024–1032.

Uzuner, O., Sibanda, T. C., Luo, Y., and Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42:13–35.

van Helden, P. (2012). Data-driven hypotheses. *EMBO reports*, 14(2):104–104.

van Leeuwen, M. E. and Harte, J. M. (2017). Violence against mental health care professionals: Prevalence, nature and consequences. *The Journal of Forensic Psychiatry & Psychology*, 28(5):581–598.

van Toledo, C. and Spruit, M. R. (2016). Adopting privacy regulations in a data warehouse — A case of the anonimity versus utility dilemma. In Fred, A. L. N., Dietz, J. L. G., Aveiro, D., Liu, K., Bernardino, J., and Filipe, J., editors, *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Porto, Portugal*, pages 234–239. SciTePress.

Vega-Pons, S. and Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372.

Velupillai, S., Dalianis, H., Hassel, M., and Nilsson, G. H. (2009). Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Venkatasubramanian, G. and Keshavan, M. S. (2016). Biomarkers in psychiatry - a critique. *Annals of Neurosciences*, 23(1):3–5.

Viljoen, J. L., Cochrane, D. M., and Jonnson, M. R. (2018). Do risk assessment tools help manage and reduce risk of violence and reoffending? a systematic review. *Law and Human Behavior*, 42(3):181–214.

Viljoen, J. L., McLachlan, K., and Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment*, 17(3):377–395.

Vleugel, A., Spruit, M., and van Daal, A. (2010). Historical data analysis through data mining from an outsourcing perspective: the three-phases model. *International Journal of Business Intelligence Research*, 1(3):42–65.

Wand, T. (2011). Investigating the evidence for the effectiveness of risk assessment in mental health care. *Issues in Mental Health Nursing*, 33(1):2–7.

Wang, Y. and Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70:287–299.

Wang, Y., Kung, L., Wang, W. Y. C., and Cegielski, C. G. (2014). An integrated big data analytics-enabled transformation model: Application to health care. In *Proceedings of Thirty Fifth International Conference on Information Systems, Auckland, New Zealand*.

Weaver, W. (1955). Translation. In *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA.

Weber, G. M., Mandl, K. D., and Kohane, I. S. (2014). Finding the missing link for big biomedical data. *JAMA*.

Weber, G. M., Murphy, S. N., McMurry, A. J., et al. (2009). The shared health research information network (SHRINE): A prototype federated query tool for clinical data repositories. *Journal of the American Medical Informatics Association*, 16(5):624–630.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12(4):e0174944.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464.

Whitson, J. R. (2013). Gaming the quantified self. *Surveillance & Society*, 11(1/2):163–176.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10):1–23.

Wilson, G., Aruliah, D. A., Brown, C. T., et al. (2014). Best practices for scientific computing. *PLoS Biology*, 12(1):e1001745.

Wu, Y., Jiang, M., Lei, J., and Xu, H. (2015). Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624–628.

Wyatt, J. (1991). Information for clinicians: Use and sources of medical knowledge. *The Lancet*, 338(8779):1368–1373.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. (2018). Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6):1–40.

Yadav, S., Ekbal, A., Saha, S., and Bhattacharyya, P. (2016). Deep learning architecture for patient data de-identification in clinical records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 32–41.

Yang, M., Wong, S. C. P., and Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, 136(5):740–767.

Yin, R. K. (2009). *Case Study Research – Design and Methods*. SAGE Publications.

Yoo, I., Alafaireet, P., Marinov, M., et al. (2011). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4):2431–2448.

Youssef, A. E. (2014). A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. *International Journal of Ambient Systems and Applications*, 2(2).

Zakim, D. and Schwab, M. (2015). Data collection as a barrier to personalized medicine. *Trends in Pharmacological Sciences*, 36(2):68–71.

Zelkowitz, M. V. and Wallace, D. R. (1998). Experimental models for validating technology. *Computer*, 31(5):23–31.

Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657. MIT Press.

Zhu, L. and Zheng, W. J. (2018). Informatics, data science, and artificial intelligence. *JAMA*, 320(11):1103.

List of publications

1. Menger, V., Spruit, M., Hagoort, K., and Scheepers, F. (2016). Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Computational and Mathematical Methods in Medicine*, 2016:1–11
2. Scheepers, F., Menger, V., and Hagoort, K. (2018). Datascience in de psychiatrie. *Tijdschrift voor psychiatrie*, 60(3):205–209
3. Menger, V., Scheepers, F., and Spruit, M. (2018a). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6):981
4. Menger, V., Scheepers, F., van Wijk, L. M., and Spruit, M. (2018b). DEDUCE: A pattern matching method for automatic de-identification of dutch medical text. *Telematics and Informatics*, 35(4):727–736
5. Hagoort, K., Menger, V., Velders, F., Deschamps, P., and Scheepers, F. E. (2018). 29.3 application of new data sources: Text mining and story banking. *Journal of the American Academy of Child & Asolescent Psychiatry*, 57(10):S314
6. Menger, V., Spruit, M., de Bruin, J., Kelder, T., and Scheepers, F. (2019a). Supporting reuse of EHR data in healthcare organizations: The CARED research infrastructure framework. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, volume 5: HEALTHINF, pages 41–50. SCITEPRESS - Science and Technology Publications
7. Menger, V., Spruit, M., van der Klift, W., and Scheepers, F. (2019b). Using cluster ensembles to identify psychiatric patient subgroups. In *Artificial Intelligence in Medicine*, pages 252–262. Springer International Publishing
8. Menger, V., Spruit, M., van Est, R., Nap, E., and Scheepers, F. (2019c). Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Network Open*, 2(7):e196709

Summary

Despite the progress that modern day health care has made in improving people's health and well-being, there are still many open questions related to disease and treatment. Combined with the ever increasing costs of health care, there is a need for new and innovative approaches, both to further expand medical knowledge and to keep health care affordable. Analyzing Electronic Health Records (EHRs) is such a potential source of innovation, since EHRs often contain information that is hidden on an aggregated level. This kind of knowledge can be made explicit through a *knowledge discovery* process, in which new and useful information is extracted from data, for example using techniques from *machine learning* and *natural language processing*. In this research, we focus on analyzing EHRs in psychiatry, the field that specializes in mental health care. Psychiatry faces some interesting challenges, making it a good example for investigating how knowledge discovery in EHRs can contribute to health care. We therefore pose the following overarching research question:

How can data from Electronic Health Records provide relevant insights for psychiatric care?

In the first three research chapters of this work, we identify key technical, organizational and ethical challenges related to knowledge discovery in EHRs, for which we subsequently propose solutions. First, we look at collaboration between *data experts*, well versed in the technical part of data analysis, and *practitioners*, who are an excellent source of *domain knowledge*. We introduce the CRISP-IDM process, where the I stands for Interactive, as a process model for collaboration based on data visualization. We show how new knowledge and hypotheses can be found using this process, most of which were not imagined beforehand. Secondly, we investigate how to design technical infrastructure, consisting of hardware and software components, that enables using EHR data for analysis. We introduce the Capable Reuse of EHR Data (CARED) framework, aiming to support health care institutions to design such infrastructure. It addresses nine important requirements, such as integrating data sources, support for collaboration and documentation, repeatability of analysis, and

privacy and security. Thirdly, we develop and validate the De-identification Method for Dutch Medical Text (DEDUCE), which aims to automatically remove information that can identify a patient from free text. It is a rule-based method that successfully removes information in categories such as person names, geographical locations, and names of health care institutions.

In the second part of this research, we focus on applying knowledge discovery techniques to EHR data to obtain new insights with potential to improve care. First we look at *violence risk assessment*, currently often performed using structured violence risk instruments or unstructured clinical judgment. We investigate whether applying machine learning techniques to clinical notes from patients' EHRs is a fruitful novel approach. After exploring which types of models, including relatively recent *deep learning* models, show promise for such a classification task, we obtain datasets of psychiatric admissions and clinical notes from EHRs of two independent psychiatric health care institutions. We use these two datasets to train models that can assess violence risk based on clinical text, and then perform a rigorous evaluation of their accuracy and generalizability. Our findings show that such models have definite potential for use in practice. Finally, we turn to identifying psychiatric patient subgroups, and investigate how *unsupervised learning* can find robust and accurate stratifications of patients. We use *cluster ensembles*, combinations of multiple clusterings, to obtain three significant clusters of adolescent patients, and assess their meaning and relation to other relevant clinical variables.

The two parts of this dissertation combined show that learning from EHRs, after addressing key challenges related to the nature of data, is a new and interesting approach with clear potential for improving psychiatric health care.

Samenvatting

Ondanks de vooruitgang die in de gezondheidszorg van vandaag de dag is geboekt in het verbeteren van de gezondheid en het welzijn van mensen, zijn er nog steeds vele open vragen gerelateerd aan ziekte en behandeling. Gecombineerd met de immer toenemende kosten van gezondheidszorg is er een noodzaak voor nieuwe en innovatieve benaderingen, zowel om medische kennis uit te breiden en om de gezondheidszorg betaalbaar te houden. Het analyseren van elektronische patiëntendossiers (EPDs) is zo'n potentiële bron van innovatie, aangezien EPDs vaak informatie bevatten die op een geaggregeerd niveau verscholen zit. Dit soort kennis kan expliciet gemaakt worden door middel van een *knowledge discovery*-proces, waarin nieuwe en bruikbare informatie wordt geëxtraheerd uit data, bijvoorbeeld met technieken uit de *machine learning* en *natural language processing*. In dit onderzoek richten we ons op het analyseren van EPDs in de psychiatrie, het veld dat zich specialiseert in geestelijke gezondheidszorg. De psychiatrie heeft te maken met enkele interessante uitdagingen, wat het tot een goed voorbeeld maakt om te onderzoeken hoe knowledge discovery in EPDs kan bijdragen aan gezondheidszorg. We stellen daarom de volgende overkoepelende onderzoeksvraag:

Hoe kan data uit elektronische patiëntendossiers relevante inzichten voor psychiatrische zorg bieden?

In de eerste drie onderzoekshoofdstukken uit dit werk identificeren we belangrijke technische, organisatorische, en ethische uitdagingen gerelateerd aan knowledge discovery in EPDs, waarvoor we vervolgens oplossingen voorstellen. Als eerste kijken we naar samenwerking tussen *data experts*, goed thuis in het technische deel van analyse, en *behandelaren*, die een uitstekende bron van *domeinkennis* zijn. We introduceren het CRISP-IDM proces, waarin de I staat voor Interactief, als een procesmodel voor samenwerking gebaseerd op data visualisatie. We laten zien hoe aan de hand van dit proces nieuwe kennis en hypothesen gevonden kunnen worden, waarvan het grootste deel van tevoren niet bedacht was. Als tweede onderzoeken we hoe een technische infrastructuur, bestaande uit hardware en software, ontworpen kan worden, die het gebruik van EPD data voor analyse mogelijk maakt. We introduceren het Ca-

pable Reuse of EHR Data (CARED) framework, met als doel het ondersteunen van zorginstellingen die zo'n infrastructuur willen ontwerpen. Het framework richt zich op negen belangrijke vereisten, zoals het integreren van databronnen, ondersteuning voor samenwerking en documentatie, herhaalbaarheid van analyse, en privacy en veiligheid. Als derde ontwikkelen en valideren we de De-identification Method for Dutch Medical Text (DEDUCE), welke als doel heeft om automatisch informatie die een patiënt kan identificeren te verwijderen uit vrije tekst. Het is een regel-gebaseerde methode, die met succes informatie uit categorieën zoals persoonsnamen, geografische locaties, en namen van zorginstellingen verwijdert.

In het tweede deel van dit onderzoek richten we ons op het toepassen van knowledge discovery technieken op data uit EPDs om nieuwe inzichten te verkrijgen met potentie voor zorgverbetering. Als eerste kijken we naar *taxatie van agressierisico*, nu vaak gedaan aan de hand van gestructureerde risicotaxatie-instrumenten of het ongestructureerde klinische oordeel. We onderzoeken of het toepassen van machine learning technieken op klinische teksten uit de EPDs van patiënten een vruchtbare nieuwe aanpak is. Nadat we hebben verkend welke soorten modellen, waaronder relatief recente *deep learning* modellen, veelbelovend zijn voor zo'n classificatietaak, verwerven we twee datasets bestaande uit psychiatrische opnames en klinische teksten uit de EPDs van twee onafhankelijke psychiatrische zorginstellingen. We gebruiken deze datasets om modellen te trainen die risicotaxatie uitvoeren op basis van klinische tekst, en voeren daarna een rigoureuze evaluatie uit van de nauwkeurigheid en generaliseerbaarheid. Onze bevindingen laten zien dat zulke modellen duidelijke potentie hebben voor gebruik in de praktijk. Tot slot richten we ons op het identificeren van subgroepen van psychiatrische patiënten, en onderzoeken hoe we door middel van *unsupervised learning* robuuste en nauwkeurige indelingen van patiënten kunnen vinden. We maken gebruik van *cluster ensembles*, combinaties van meerdere clusteringen, om drie significante clusters van adolescente patiënten te vinden, en stellen hun betekenis en relatie met andere relevante klinische variabelen vast.

De twee delen van dit proefschrift samen laten zien dat leren van EPDs, nadat belangrijke uitdagingen met betrekking tot het soort data zijn geadresseerd, een nieuwe en interessante benadering is, met duidelijke potentie voor het verbeteren van psychiatrische zorg.

Curriculum Vitae

Vincent Menger was born on July 16th, 1989, in Franeker, the Netherlands. He attended Utrecht University between 2008 and 2015, where he obtained a Bachelor's degree in Cognitive Artificial Intelligence in 2012, and a Master's degree (*cum laude*) in Technical Artificial Intelligence in 2015. He wrote his Master's thesis on the *pattern explosion*, a phenomenon in Pattern Set Mining where the number of patterns eventually exceeds the number of data points they are obtained from.

In 2015, he started his research as a PhD candidate at the Department of Information and Computing Sciences of Utrecht University and the Department of Psychiatry of the University Medical Center Utrecht. During his time as a PhD researcher, he presented his work at various national and international scientific meetings, including a masterclass of prof. Thomas Insel and the 62nd Annual Meeting of the American Association of Child and Adolescent Psychiatry. One of his research contributions on violence risk assessment was covered in an article by the national newspaper NRC Handelsblad. His teaching contributions included giving lectures and leading lab sessions for various courses including Data Analytics, Data Science & Society, and Scientific Research Methods, as well as supervision of graduation projects. Simultaneous to doing his research, he worked as a data scientist at the Department of Psychiatry, working closely with practitioners in uncovering new information from patient records, always with the ultimate goal to improve people's mental health in mind.