

Doctoral thesis

Doctoral theses at NTNU, 2023:448

Thomas Brox Røst

Enabling Data-Driven Decision Support in Healthcare

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Department of Computer Science



Norwegian University of
Science and Technology

Thomas Brox Røst

Enabling Data-Driven Decision Support in Healthcare

Thesis for the Degree of Philosophiae Doctor

Trondheim, December 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science



Norwegian University of
Science and Technology

NTNU

Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Computer Science

© Thomas Brox Røst

ISBN 978-82-326-7600-2 (printed ver.)
ISBN 978-82-326-7599-9 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:448

Printed by NTNU Grafisk senter

The real problem is not whether machines think but whether men do.

- B. F. SKINNER

Abstract

Healthcare professionals make decisions every day. These decisions can have a profound impact on those they concern, such as a decision on how to treat a patient. Patients and their families can experience prolonged illness or death if the decision results in an adverse event; the treating healthcare organization and society as a whole will suffer the additional clinical, economic and social impact. Getting decisions right is important for safe provisioning of healthcare.

This thesis discusses ways of helping healthcare professionals make better decisions through targeted use of healthcare data and knowledge. We specifically consider these three research questions:

- RQ1** How can health-related data and knowledge aid decision-making?
- RQ2** How can healthcare professionals contribute to implementation of data-driven decision-making?
- RQ3** How can patient trajectory data contribute to decision-making in healthcare?

The five included publications cover various aspects of data- and knowledge-driven decision-making in the healthcare domain. The application scope ranges from decision support in the electronic health record to automated clinical knowledge classification. Results from the publications are discussed in light of the research questions. Finally, we consider how recent machine learning advances may affect decision support in healthcare, both in terms of possibilities and obstacles.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfillment of the requirements for the degree of Philosophiae Doctor. This doctoral work was performed at the Department of Computer Science, NTNU, Trondheim, Norway under the supervision of Associate Professor Øystein Nytrø and earlier co-supervision of Professor Emeritus Anders Grimsmo and Professor Torbjørn Nordgård.

Funding was provided through three different Research Council of Norway (RCN) projects: EVICARE (project number 193022), BIGMED (project number 259055), and IDDEAS (project number 269117).

Thomas Brox Røst

Trondheim, Norway, September 2023

Acknowledgements

This thesis would not have existed without the contributions of all my colleagues and co-authors over the years. Special thanks go to my advisor, Øystein Nytrø, for making it possible to finish my work, for continuously reminding me to do so, and for invaluable feedback over the years. Thanks also go to the people at the department, faculty, and university administration for allowing me to defend this thesis after having spent way more time on it than I should.

Thank you to Yvonne for all encouragement and care during the writing process. It wouldn't have happened without you.

Contents

ABSTRACT **V**

PREFACE **VII**

ACKNOWLEDGEMENT **IX**

LIST OF FIGURES **XV**

NOMENCLATURE **XVII**

PART I RESEARCH OVERVIEW AND SUMMARY 1

CHAPTER 1 INTRODUCTION 3

- 1.1 Scope and Definitions **4**
- 1.2 Electronic Health Records and Secondary Use **4**
 - 1.2.1 Unstructured Data **5**
 - 1.2.2 Data Quality **5**
 - 1.2.3 Ethics and Privacy **6**
 - 1.2.4 Bias **7**
 - 1.2.5 Temporal Representations **7**
- 1.3 Clinical Decision Support **8**
 - 1.3.1 Adoption of CDSS **8**
 - 1.3.2 Explainable Decision Support **9**
- 1.4 Evidence Mapping and Systematic Reviews **10**
- 1.5 Background and Motivation **11**
- 1.6 Research Questions **12**

1.7	Publications and Contributions	13
1.8	Ethical Approval and Funding	14
1.9	Thesis Outline	14
1.10	Other Publications	14
1.11	References	15
CHAPTER 2	RESULTS	23
2.1	Paper A: Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-Initiated Project	23
2.1.1	Paper Summary	24
2.2	Paper B: Identifying Catheter-Related Events Through Sentence Classification	25
2.2.1	Paper Summary	26
2.3	Paper C: Using Neural Networks to Support High-Quality Evidence Mapping	26
2.3.1	Paper Summary	27
2.4	Paper D: Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health	27
2.4.1	Paper Summary	28
2.5	Paper E: Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development	29
2.5.1	Paper Summary	31
2.6	References	31
CHAPTER 3	DISCUSSION	35
3.1	Paper A: Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-Initiated Project	35
3.2	Paper B: Identifying Catheter-Related Events Through Sentence Classification	36

3.3	Paper C: Using Neural Networks to Support High-Quality Evidence Mapping	37
3.4	Paper D: Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health	37
3.5	Paper E: Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development	38
3.6	References	38
CHAPTER 4	CONCLUSION	39
4.1	Deep Learning and Foundation Models	39
4.2	Interface Expectations	40
4.3	Obstacles to AI in Medicine	40
4.4	Human-AI Collaboration	41
4.5	The Road Ahead	42
4.6	References	43
<hr/>		
PART II PUBLICATIONS		47
PAPER A: DEVELOPMENT OF A MEDICATION RECONCILIATION TOOL FOR NORWEGIAN PRIMARY CARE EPR SYSTEMS: EXPERIENCES FROM A USER-INITIATED PROJECT		A-1
PAPER B: IDENTIFYING CATHETER-RELATED EVENTS THROUGH SENTENCE CLASSIFICATION		B-1
PAPER C: USING NEURAL NETWORKS TO SUPPORT HIGH-QUALITY EVIDENCE MAPPING		C-1
PAPER D: LOCAL, EARLY, AND PRECISE: DESIGNING A CLINICAL DECISION SUPPORT SYSTEM FOR CHILD AND ADOLESCENT MENTAL HEALTH		D-1

PAPER E: USABILITY OF THE IDDEAS PROTOTYPE IN CHILD AND ADO-
LESCENT MENTAL HEALTH SERVICES: A QUALITATIVE STUDY
FOR CLINICAL DECISION SUPPORT SYSTEM DEVELOPMENT **E-1**

List of Figures

CHAPTER 1

CHAPTER 2

- 2.1 Screenshot of the medication reconciliation tool in use within the Infodoc EPR system. 24
- 2.2 The IDDEAS clinical decision support model. 28
- 2.3 Individualized Digital Decision Assist System prototype software screenshot. 30

CHAPTER 3

CHAPTER 4

Nomenclature

ACRONYMS

ADHD	Attention deficit hyper-activity disorder	GMAI	Generalist medical AI
AI	Artificial intelligence	GP	General practice
ATC classification	Anatomical Therapeutic Chemical classification	HCP	Healthcare professional
BSI	Blood-stream infection	HIPAA	Health Insurance Portability and Accountability Act
CAMHS	Child and adolescent mental health services	ICT	Information and communication technology
CDSS	Clinical decision support system	LLM	Large language model
CVC	Central venous catheter	ML	Machine learning
EHR	Electronic health record	NIPH	Norwegian Institute of Public Health
EPR	Electronic patient record	NLP	Natural language processing
EU	European Union	RCN	Research Council of Norway
GDPR	General Data Protection Regulations		

Part I

RESEARCH OVERVIEW AND SUMMARY

CHAPTER 1

Introduction

Decision-making is a key part of healthcare processes. When decisions turn out right, few second thoughts are given; when they turn out wrong we seek answers as to why. Making wrong decisions during patient treatment can have a negative impact on multiple levels, in particular if the outcome of the decision is an adverse event¹. The direct impact is felt by the patient and family members, be it through psychological damage, physical harm, or even death. The *second victims*, to use a term introduced by Wu, are the healthcare professionals responsible for the decision²; estimates are that between 10 and 50 percent of clinicians involved with adverse events have suffered some form of emotional impact. Ultimately, adverse events that affect individuals have a negative effect on wider society as well. Wrong decisions can also have a negative impact without being directly harmful. Other possible outcomes are unhelpful, unnecessary, or sub-optimal care. This can have effects such as unnecessary prolongation of the episode of care and inefficient use of healthcare resources.

Regardless of the impact of wrong decisions, we prefer avoiding them. If an option existed to spend less time and effort on decisions without sacrificing quality, precision and outcome, most of us would take it. This thesis looks at ways of making technology support decisions in healthcare processes in a way that hopefully makes them more precise, more efficient, and more beneficial to those impacted by them. We consider both the data and knowledge that feeds the decision-making process as well as the healthcare professionals that are responsible for the outcome of that process. The publications within highlight various healthcare application domains where decision support can make a difference. We also give an overview of how recent technological advances within machine learning and artificial intelligence may present new opportunities towards making clinical decision support work in practice—but also how some key obstacles remain.

1.1 SCOPE AND DEFINITIONS

The title of this thesis, *Enabling Data-Driven Decision Support in Healthcare*, covers the main subject areas in the included publications and the discussion.

Starting at the end, *healthcare* limits the scope to the functions performed by healthcare professionals (HCPs) making decisions regarding patient treatment and evidence mapping in evidence-based medicine. With *decision-support* the focus is on application areas that support health professionals in making decisions. Furthermore, *data-driven* means that the decision support relies on analysis and processing of *data* from the targeted healthcare domain that has relevance for the decision-making process. While not explicitly mentioned, *information* and *knowledge* also feeds into this process. In the context of this thesis *data* implies clinical texts, *information* text in medical research publications, and *knowledge* structured clinical guidelines; the latter as executable or interpretable clinical guidelines and recommendations. Finally, *enabling* is either the realization of software systems that provide the indented decision support functionality or establishing some fundamentals for future implementations. In this thesis this implies understanding the design requirements and contextual needs of such systems. It can also mean production deployments of decision-support software applications in the software engineering sense.

The research within is fundamentally applied research. The data-driven research approaches are therefore more about applying known data-processing and analysis methods on novel research questions rather than enhancing the methods themselves. Also, some publications are more knowledge- than data-driven; this is reflected in the research questions below. In terms of the data used we limit ourselves to clinical record data and research publication data. The main focus is on unstructured text data, with machine learning (ML) and natural language processing (NLP) methods being the primary methodological tools.

The terms *machine learning* (ML) and *natural language processing* (NLP) are used interchangeably to refer to techniques for automated sense-making from unstructured clinical text, while *artificial intelligence* (AI) is used to denote applications that are enabled by ML without necessarily being intelligent.

1.2 ELECTRONIC HEALTH RECORDS AND SECONDARY USE

The electronic health record (EHR) is the primary tool to store, organize, and communicate information about patient treatment³. As technologies for generating various forms of patient- and care-related data have become more prevalent the amount of available healthcare data has grown significantly. Our

conventional approaches for health data management are, as described by Fang et al.⁴, no longer able to cope with the high-volume and -velocity influx of data that we are seeing today. This has led to increased focus on the potential for secondary use of such data for both clinical research⁵ and healthcare practice improvement purposes⁶, with the former being of particular interest following the recent COVID-19 pandemic^{7,8}. Data from EHRs gives a glimpse into medical care and outcomes for a diverse population of patients and may help reduce research costs, increase patient-centered research, and boost medical discoveries⁹.

1.2.1 *Unstructured Data*

EHRs contain a wide and rich range of patient data in both structured, semi-structured, and unstructured format, with the latter typically making up the bulk of the data volume in the form of free-text in e.g. clinical notes and discharge summaries. While textual data contains rich information about patient treatment, extracting knowledge from such data is not a trivial task¹⁰. First of all, the primary purpose of collecting the data is to document patient treatment and not to support secondary research use. This means that the data available to us may have varying degrees of relevance for our research objectives. Moreover, without access to the original author of the data and their intentions a layer of interpretation will always have to be added. To further complicate the problem, natural language will be ambiguous, non-standardized, contain spelling errors, be in violation of grammar rules and have other idiosyncrasies⁶.

Natural language processing (NLP) and machine learning (ML) techniques have been used to make sense of unstructured text in EHRs^{11,12} for a wide variety of research purposes¹³. Recent advances in machine learning methods along with the growing availability of data and computing power has led to steady advancements in the state of the art and increased belief and interest in the viability of such methods for enabling secondary use of EHR and medical data in general¹⁴.

1.2.2 *Data Quality*

In their review of methods and dimensions of data quality assessment in EHRs, Weiskopf and Weng identified the following five dimensions of data quality⁹:

- **Completeness:** Is the truth about a patient present in the EHR?
- **Correctness:** Is an element that is present in the EHR true?

- **Concordance:** Is there agreement between elements in the EHR, or between the EHR and another data source?
- **Plausibility:** Does an element in the EHR make sense in light of other knowledge about what that element is measuring?
- **Currency:** Is an element in the EHR a relevant representation of the patient state at a given point in time?

High-quality EHR data that are suitable for clinical research, quality improvement and public health purposes rarely fulfill all of these quality dimensions¹⁵. A fundamental limitation is that data will only be recorded during healthcare episodes, i.e., when the patient has an illness. Information about conditions associated with good health is likely absent. As for the data that is available, a lot of it will be implicit rather than explicit, such as e.g. negative findings. Data fragmentation as a result of the patient moving between different institutions and lack of interoperability between patient information systems is also an issue. A practical effect seen in our research is that the actual start and end points of an episode of care often have to be deduced from e.g. admission and discharge notes. As mentioned, the bulk of the data will normally be in the form of narrative notes, which further complicates the task of converting the required data into a usable structured format. Data may be structured and coded according to clinical concept models, taxonomies and classifications schemes. Still, the use of these may vary across healthcare organization and drift over time even within a single institution.

There are approaches towards making EHR data more applicable for secondary use. A typical process is to apply phenotyping or feature extraction on raw EHR data to transform it into clinically relevant features and then use these features for research purposes. However, at scale this can be a cumbersome and inexact effort, requiring, as described by Hripcsak and Albers, "detective work and alchemy to get golden phenotypes from base data"¹⁵.

1.2.3 *Ethics and Privacy*

While the potential benefits of using EHR data for health service improvements are huge, both for individual patients and the public at large, this must be done in an ethical and privacy-preserving manner^{16,17}. This involves ensuring that privacy and data security measures are in place, having informed consent and establishing data ownership, but also the provisioning of ethics training and accountability for ethical treatment of sensitive patient data. Risks involved with the use of EHR data should be minimized as much as possible, primarily through traditional safeguards but also by considering strategies such as

data de-identification^{18,19} and the use of synthetic patient data²⁰. However, even such approaches are not without flaws. While both the European Union's General Data Protection Regulations (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule do not apply to technically anonymous data, patients can still potentially be re-identified from such data sets, as noted by Rocher et al.²¹ The use of synthetic data has a lot of potential as a privacy-preserving tool but as techniques for creating believable synthetic data become more sophisticated, the risk for malicious or accidental influence on public perception of clinical ground truth increases. Chen et al. describes how this can affect the AI algorithms that increasingly are part of decisions with consequences for our lives²². Still, synthetic data will not necessarily be less suitable for its intended purposes than the likely flawed data it is based on.

1.2.4 *Bias*

On the most basic level, clinician workload, user interface design issues, and other data entry factors means that EHR data is often fragmented, incomplete or simply erroneous^{23–25}. When investigating the effect of data completeness and validity on prescription decisions in an intensive care unit, Kramer et al. found that data quality can have a direct impact on decisions without practitioners being aware of this²⁶. Given complex disease histories and the presence of comorbidity, confounding bias is another factor that may impact the validity of the secondary use purpose^{27,28}. Data sets can also carry forward the biases we bring to it. Caliskan et al. showed that imprints of historical bias against race and gender could be recovered from a machine learning model trained on a standard World Wide Web text corpora²⁹. Moreover, since there will always be proportionally less data about minorities, the use of such imbalanced data sets in machine learning classifiers will mean that they are more likely to perform worse for underrepresented groups, as described by Schönberger³⁰, or for rare conditions. Finally, the focus on conditions present for illness rather than health is a fundamental limitation and source of bias in any data collected and documented as part of the provisioning of healthcare.

1.2.5 *Temporal Representations*

While the temporal data in the EHR can be used for insight into e.g. disease progression and treatment trajectories, transforming this data into actionable knowledge is not straightforward³¹. In the words of Hripcsak and Albers¹⁵, the "EHR is not a direct reflection of the patient and physiology, but a reflection of

the recording process inherent in healthcare with noise and feedback loops"—or, in other words, the patient's interaction with the healthcare system. As an example, the date of a diagnosis code tells us when a physician made the diagnosis but not when the patient developed the illness. Furthermore, the diagnosis code itself may not only be a result of observations of and tests performed on the patient but also of the reimbursement policies in place at the time of recording. Using EHR data without taking the context in which it was produced into account can lead to bias that diminishes its value for answering research questions. A good example of this can be found in a study by Agniel et al. where they compared the predictive value for survival of laboratory test result values (i.e., patient pathophysiology) and the timing of when the test was ordered (i.e., the underlying healthcare process)³². It turned out that for most tests the hour of the day, the day of the week and the time between tests was more predictive of survival than the test results themselves. While this warns against naive use of EHR data, their findings also showed that understanding and explicitly modeling the healthcare process dimension could to some extent increase the predictive value of the data.

1.3 CLINICAL DECISION SUPPORT

Clinical decision support systems (CDSS) is traditionally defined as software that aids clinical decision-making by matching the characteristics of individual patients to a clinical knowledge base, this in order to provide patient-specific assessments and recommendations to the clinician or patient³³. They are typically used at the point-of-care to enhance the knowledge of the clinician through relevant suggestions. A more recent approach is to utilize data- and observation-driven approaches as well. One way of classifying CDSSs is as being either knowledge- or non-knowledge-based³⁴. The latter approach relies on the combination of data sources and machine learning rather than a knowledge base representing expert clinical domain knowledge.

1.3.1 Adoption of CDSS

In their article on CDSS implementation successes and failures, Greenes et al.³⁵ remind us that in spite of a history going back more than five decades, widespread adoption in clinical practice has still not been achieved. Improved practical outcomes are far from given, not only on patient treatment^{36,37} but also in terms of economic, workload, and efficiency outcomes³⁸. Several studies have attempted to figure out success factors for making CDSS work^{35,36,39,40}. Common problems include disruptions to established clinician workflows⁴¹,

alert fatigue⁴², over-reliance on automation⁴³, lack of technical skills among users⁴⁴, CDSS maintenance and upkeep⁴⁵, and lack of interoperability and integration between CDSSs and EHRs³⁴. Even the experience level of the clinician comes into play. In an analysis by Dowding et al.⁴⁶, practitioners with more experiential knowledge were found to be more likely to override CDSS suggestions. At the time of writing, machine learning systems for automated interpretation of medical images within radiology are approaching human-level performance and are being approved for clinical use⁴⁷. Still, figuring out how to interface between the decision-making clinician and the CDSS remains a challenge. In a recent working paper by Agarwal et al.⁴⁸ studying the effectiveness of human-AI collaboration in radiology it was found that the radiologists tended to under-weigh the AI's information versus their own and that AI predictions did not by themselves increase diagnostic quality. As it turns out, for the evaluated system the optimal solution was to assign cases either to humans or to the AI, rather than having the AI assist the human. In other studies the opposite effect is seen, in that the AI and human working together is the optimal solution^{33,49}.

1.3.2 *Explainable Decision Support*

The increased focus on non-knowledge based CDSS under the "AI in medicine" umbrella term comes with its own set of challenges, including opaque, black-box machine learning model logic and issues around data availability and ethical usage^{17,37}. He et al. argue that transparency and interpretability is a fundamental requirement for machine learning systems in healthcare, not only for understanding how a decision or prediction was made but also for helping us uncover new clinical insights⁵⁰. This also applies to the previously mentioned concerns about bias against minorities in clinical data sets. According to Char et al.¹⁷, another risk from opaqueness is that of direct manipulation to boost e.g. certain drugs and tests in the given suggestions. Indeed, for citizens of the EU there is an ongoing debate on to what extent the GDPR gives you the right to receive an explanation for algorithmic decisions that may affect you^{51,52}. For these reasons there is increased focus on bridging the gap between ML models and healthcare professionals through explainable machine learning. Explainability does, however, come with its own costs. A recent study by Wysocki et al. describing an evaluation framework for model explainability found that while explanations in general were perceived positively, they also often increased the cognitive effort for the healthcare professionals without improving the participant's understanding of the machine learning model⁵³. Increased risk for confirmation bias, i.e. when the recommendation agrees with the user's initial

decision, was also seen. On the positive side, explanations were also seen to reduce automation bias and help participants acquire new domain knowledge and reach quicker decisions.

1.4 EVIDENCE MAPPING AND SYSTEMATIC REVIEWS

An evidence map gives a broad overview of the volume, nature, and characteristics of a research field while a systematic review focuses on a single clinical question⁵⁴. Evidence maps can be used to complement systematic reviews by showing how research evidence maps to various populations and contexts and making it evident where there are gaps and potential for additional research⁵⁵. Even before the recent global pandemic the doubling time of medical knowledge had reduced dramatically: from 50 years in 1950 to 7 years in 1980 and 3.5 years in 2010⁵⁶. The pandemic further increased the rate of medical knowledge dissemination but accompanied by concerns about quality and trustworthiness. Glasziou et al.⁵⁷ remind us that prior to COVID-19 an estimated 85% of research was wasted due to factors such as poor study design and poor reporting of results. This problem only seemed to be amplified with the rush of COVID-19-related research. "Living" evidence synthesis ecosystems have been proposed as a tool to help us navigate this vast volume of information and, increasingly, data, making sure that the right information is available for clinical decision-making⁵⁸.

The systematic review methodology is rigorous, resource-intensive, and time-consuming⁵⁸. Thomas et al. claim that automation-supported workflows, where tools such as machine learning and natural language processing can support human reviewers, is a possible way to reduce the associated workload⁵⁹. There is in general room for improvement in the systematic review process, be it through use of technology or otherwise⁶⁰. Even where systematic reviews exist, they are not always used to their full potential. In an assessment of the use of systematic methods for guideline evidence synthesis by Lunny et al., only half of the included studies took a systematic approach to evidence synthesis⁶¹. Several systems that support automation of systematic reviews already exist but a recent scoping review by Khalil et al. found that while many have potential they still have come with limitations⁶². Techniques for abstract screening are mature enough to be useful but data extraction is still an active research area. In many cases the use of active learning, where continuous improvement and input is given through human involvement in the review workflow, tends to be associated with improved performance when compared with full automation. Also, similar to CDSSs, the importance of moving from research prototypes to professionally maintained and deployed platforms is stressed.

Systematic reviews are a necessary requirement for *evidence-based medicine*, which Sackett describes as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."⁶³ In practice this means making the best and most relevant clinical evidence available to support the clinician's individual expertise and experience. This evidence can then be used to create *clinical guidelines*, which are used to help practitioners make informed decisions about appropriate care for specific clinical situations⁶⁴. Guidelines are often used in the implementation of clinical decision support, but this is not always straightforward. One problem, as described by Sittig et al. in their review of challenges in clinical decision support⁶⁵, is that guidelines often fail to address the complexities of co-morbidity and polypharmacy, which will often be the case with e.g. elderly patients. This may be part of the explanation why clinical guidelines are often less used in clinical practice than they should be.

1.5 BACKGROUND AND MOTIVATION

The original objective of this PhD was to investigate the use natural language processing (NLP) and machine learning (ML) methodologies applied to clinical text for information extraction purposes, with particular focus on its application to patient histories and how the broader context of a finding in a clinical note could contribute to better understanding it. The target domain was clinical text from general practice; the motivation being the early digitization of general practice in Norway and thus access to long-term historical clinical text data.

Over time this objective became both more open-ended and more specific. The original focus on data from general practice grew to also include hospital and research publication data. While the use of language processing and machine learning methodologies remained, the scope also broadened to the use of clinical guidelines. The publications cover a wide range of application domains, ranging from NLP-based medication reconciliation in general practice to guideline-driven clinical decision support systems for specialist mental health care. In some publications the use of patient histories is still a factor; in others this is not the case.

The initial objective was exploratory in the sense that given enough patients with enough data there was probably something interesting to be found. In the work that ensued and the included publications there ended up being a stronger sense of purpose as to *why* the application of these methodologies could make a difference. In spite of wide-ranging application domains a shared common thread is the automated use of data and knowledge in order to support *decision-making* for healthcare professionals.

We consider the term decision-making to include both the established concepts of clinical decision and process support, but also the use of methods that help us apply structure to an unstructured data set in order to learn more about e.g. a body of patients or medical practices at a hospital ward so that this may influence our decisions. Patients may not even be involved; in one publication the purpose is to support knowledge-gathering and systematic review of medical research.

Another key theme in the included publications is the involvement of health-care professionals. The combination of multiple applied application domains brought forward a focus on what it takes to make such technologies work in practice, ranging from technological and legal constraints to understanding the operating context and the needs of users and stakeholders for treating patients in an effective and precise manner. Some of the work was used in a real-life healthcare setting on actual patient data. This gave additional perspective on the complexities of making data-driven decision-support technology work on sensitive patient data in a healthcare treatment and provisioning setting. Beyond the technological and legal challenges, a key takeaway is that the end user is the ultimate arbiter and decision-maker and that the role of the technology is to aid with making the best and most precise decision given the available information.

1.6 RESEARCH QUESTIONS

Given the motivation above and the thesis title, *Enabling Data-Driven Decision Support in Healthcare*, we have attempted to summarize the research objective through the following research questions:

RQ1 How can health-related data and knowledge aid decision-making?

RQ2 How can healthcare professionals contribute to implementation of data-driven decision-making?

RQ3 How can patient trajectory data contribute to decision-making in healthcare?

The questions highlight factors that contribute to the design and implementation of a CDSS: the data and knowledge that goes into the system (RQ1), the necessary contributions from those that will eventually use it (RQ2), and the representation of the patient whose treatment is affected by the CDSS in the form of patient trajectory data (RQ3).

The following section will outline how the included papers contribute to these research questions.

1.7 PUBLICATIONS AND CONTRIBUTIONS

The five publications included in this thesis can be summarized as follows:

Paper A - Medication Reconciliation: Extraction of medication events from hospital discharge summaries in order to provide process support for medication reconciliation in the general practice EPR⁶⁶.

Paper B - Central Venous Catheter Prevalence: Automated extraction of central venous catheter use from episodes of care in order to learn more about catheter use and associated adverse events⁶⁷.

Paper C - COVID-19 Research Evidence Mapping: Automated classification of COVID 19-related research publications to support a manual systematic evidence mapping process⁶⁸.

Paper D - Designing CDSS for ADHD: Designing clinical decision support systems to support treatment of ADHD in child and adolescent mental health services⁶⁹.

Paper E - Usability of CDSS for ADHD: Examining the needs of child and adolescent psychiatrists and clinical psychologists from a clinical decision support application⁷⁰.

The publications and how they address the research questions are described in detail in the Results section (Chapter 2). The list below gives an overview of the relationship between the publications and the research questions.

Paper A - Medication Reconciliation: RQ1 (applied NLP on discharge notes for automated medication reconciliation support), RQ2 (initiated by health professionals and close involvement in development and maintenance phases).

Paper B - Central Venous Catheter Prevalence: RQ1 (detection of central venous catheter use in clinical notes), RQ2 (participation in annotation process), RQ3 (method applied on full patient trajectories).

Paper C - COVID-19 Research Evidence Mapping: RQ1 (ML used for automated classification of research publications), RQ2 (coding and annotation done by stakeholders).

Paper D - Designing CDSS for ADHD: RQ1 (reflections on the role of data and knowledge of stakeholders in the treatment process), RQ3

(considerations regarding the long-term nature of CDSS for chronic or long-term illness).

Paper E - Usability of CDSS for ADHD: RQ1 (information requirements for the clinician), RQ2 (knowledge about clinician preferences and workflow in an ADHD diagnosis setting).

1.8 ETHICAL APPROVAL AND FUNDING

The process for ethical approval is described in each publication where this was necessary. The research covered by this thesis was funded through three different Research Council of Norway (RCN) projects: EVICARE (project number 193022)⁷¹, BIGMED (project number 259055)⁷², and IDDEAS (project number 269117)⁷³.

1.9 THESIS OUTLINE

The thesis is divided into two parts. The overview and summary part (Part I) starts of by introducing the research questions and the overall objective of the thesis in this chapter. It also provides definitions, scope, and necessary background material. Chapter 2 presents the results of each publication and their relevance to the research questions. Chapter 3 discusses the results in terms of the research questions and makes some observations that did not make it into the publications. Finally, Chapter 4 gives an overview of recent research developments and highlights some opportunities and challenges.

Part II contains all included publications.

1.10 OTHER PUBLICATIONS

The following publications are not included in the thesis but are still relevant for the research questions:

- *Lessons from Developing an Annotated Corpus of Patient Histories* (2008). Thomas Brox Røst, Ola Huseth, Øystein Nytrø, Anders Grimsmo. J. Comput. Sci. Eng. 2 (2), 162-179⁷⁴.
- *Comparing medical code usage with the compression-based dissimilarity measure* (2007). Thomas Brox Røst, Ole Edsberg, Anders Grimsmo, Øystein Nytrø. Studies in health technology and informatics 129 (1), 684⁷⁵.

- *Classifying encounter notes in the primary care patient record* (2006). Thomas Brox Røst, Øystein Nytrø, Anders Grimsmo. Proceedings of the 3rd International Workshop on Text-based Information Retrieval⁷⁶.

They are mentioned here but will not be discussed further.

1.11 REFERENCES

- [1] Linda T. Kohn, Janet M. Corrigan, and Molla S. Donaldson, editors. *To Err is Human: Building a Safer Health System*. Washington (DC) (2000). ISBN 0-309-06837-1. doi: 10.17226/9728. Book Title: *To Err is Human: Building a Safer Health System*. Cited on page/s 3.
- [2] Albert W. Wu, Jo Shapiro, Reema Harrison, Susan D. Scott, Cheryl Connors, Linda Kenney, and Kris Vanhaecht. The Impact of Adverse Events on Clinicians: What's in a Name? *Journal of patient safety* **16** (1), 65–72 (March 2020). ISSN 1549-8425 1549-8417. doi: 10.1097/PTS.000000000000256. Place: United States. Cited on page/s 3.
- [3] Lawrence L. Weed. Medical Records That Guide and Teach. *New England Journal of Medicine* **278** (11), 593–600 (1968). doi: 10.1056/NEJM196803142781105. URL <https://doi.org/10.1056/NEJM196803142781105>. Cited on page/s 4.
- [4] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. Computational Health Informatics in the Big Data Age: A Survey. *ACM Comput. Surv.* **49** (1) (June 2016). ISSN 0360-0300. doi: 10.1145/2932707. URL <https://doi.org/10.1145/2932707>. Place: New York, NY, USA Publisher: Association for Computing Machinery. Cited on page/s 5.
- [5] Ariel Beresniak, *et al.* Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. *Contemporary Clinical Trials* **46**, 85–91 (January 2016). ISSN 1551-7144. doi: 10.1016/j.cct.2015.11.011. URL <https://www.sciencedirect.com/science/article/pii/S1551714415301221>. Cited on page/s 5.
- [6] Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Comput. Surv.* **55** (2) (January 2022). ISSN 0360-0300. doi: 10.1145/3490234. URL <https://doi.org/10.1145/3490234>. Place: New York, NY, USA Publisher: Association for Computing Machinery. Cited on page/s 5.
- [7] Arianna Dagliati, Alberto Malovini, Valentina Tibollo, and Riccardo Bellazzi. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Briefings in Bioinformatics* **22** (2), 812–822 (March 2021). ISSN 1477-4054. doi: 10.1093/bib/bbaa418. URL <https://doi.org/10.1093/bib/bbaa418>. Cited on page/s 5.
- [8] Hua Xu, David L. Buckneridge, Fei Wang, and Peter Tarczy-Hornock. Novel informatics approaches to COVID-19 Research: From methods to applications. *Journal of Biomedical Informatics* **129**, 104028 (May 2022). ISSN 1532-0464. doi: 10.1016/j.jbi.2022.104028. URL <https://www.sciencedirect.com/science/article/pii/S1532046422000442>. Cited on page/s 5.
- [9] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20** (1), 144–151 (January 2013). ISSN 1067-5027.

- doi: 10.1136/amiajnl-2011-000681. URL <https://doi.org/10.1136/amiajnl-2011-000681>. Cited on page/s 5.
- [10] Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtlielsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics* **13** (6), e1549 (November 2021). ISSN 1939-5108. doi: 10.1002/wics.1549. URL <https://doi.org/10.1002/wics.1549>. Publisher: John Wiley & Sons, Ltd. Cited on page/s 5.
- [11] Yanshan Wang, *et al.* Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* **77**, 34–49 (January 2018). ISSN 1532-0464. doi: 10.1016/j.jbi.2017.11.011. URL <https://www.sciencedirect.com/science/article/pii/S1532046417302563>. Cited on page/s 5.
- [12] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics* **73**, 14–29 (September 2017). ISSN 1532-0464. doi: 10.1016/j.jbi.2017.07.012. URL <https://www.sciencedirect.com/science/article/pii/S1532046417301685>. Cited on page/s 5.
- [13] Irena Spasic and Goran Nenadic. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform* **8** (3), e17984 (March 2020). ISSN 2291-9694. doi: 10.2196/17984. URL <http://www.ncbi.nlm.nih.gov/pubmed/32229465>. Cited on page/s 5.
- [14] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission (2020). URL <https://arxiv.org/abs/1904.05342>. Cited on page/s 5.
- [15] G. Hripcsak and D. J. Albers. Next-Generation Phenotyping of Electronic Health Records. *Journal of the American Medical Informatics Association* **20** (1), 117–121 (2013). doi: 10.1136/amiajnl-2012-001145. URL <https://doi.org/10.1136/amiajnl-2012-001145>. Cited on page/s 6, 7.
- [16] Lisa M. Lee. Ethics and subsequent use of electronic health record data. *Journal of Biomedical Informatics* **71**, 143–146 (2017). ISSN 1532-0464. doi: <https://doi.org/10.1016/j.jbi.2017.05.022>. URL <https://www.sciencedirect.com/science/article/pii/S1532046417301211>. Cited on page/s 6.
- [17] Danton S. Char, Nigam H. Shah, and David Magnus. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England journal of medicine* **378** (11), 981–983 (March 2018). ISSN 1533-4406 0028-4793. doi: 10.1056/NEJMp1714229. Place: United States. Cited on page/s 6, 9.
- [18] Clete A. Kushida, Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care* **50 Suppl** (Suppl), S82–101 (July 2012). ISSN 1537-1948 0025-7079. doi: 10.1097/MLR.0b013e3182585355. Place: United States. Cited on page/s 7.
- [19] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. Deep Learning Architecture for Patient Data De-identification in Clinical Records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)* pages 32–41 Osaka, Japan (December 2016). The COLING 2016 Organizing Committee. URL <https://aclanthology.org/W16-4206>. Cited on page/s 7.
- [20] Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health* **2** (1), e0000082 (January 2023). doi: 10.1371/journal.pdig.0000082. URL <https://doi.org/10.1371/journal.pdig.0000082>. Publisher: Public Library of Science. Cited on page/s 7.
- [21] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success

- of re-identifications in incomplete datasets using generative models. *Nature Communications* **10** (1), 3069 (July 2019). ISSN 2041-1723. doi: 10.1038/s41467-019-10933-3. URL <https://doi.org/10.1038/s41467-019-10933-3>. Cited on page/s 7.
- [22] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* **5** (6), 493–497 (June 2021). ISSN 2157-846X. doi: 10.1038/s41551-021-00751-8. URL <https://doi.org/10.1038/s41551-021-00751-8>. Cited on page/s 7.
- [23] Sharona Hoffman and Andy Podgurski. Big Bad Data: Law, Public Health, and Biomedical Databases. *Journal of Law, Medicine & Ethics* **41** (S1), 56–60 (2013). doi: 10.1111/jlme.12040. URL <https://journals.sagepub.com/doi/10.1111/jlme.12040>. Publisher: Cambridge University Press. Cited on page/s 7.
- [24] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on translational bioinformatics* **2010**, 1–5 (March 2010). ISSN 2153-6430. Place: United States. Cited on page/s 7.
- [25] Gaurav Jetley and He Zhang. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. *Decision Support Systems* **126**, 113137 (2019). ISSN 0167-9236. doi: 10.1016/j.dss.2019.113137. URL <https://www.sciencedirect.com/science/article/pii/S0167923619301666>. Cited on page/s 7.
- [26] Oren Kramer, Adir Even, Idit Matot, Yohai Steinberg, and Yuval Bitan. The impact of data quality defects on clinical decision-making in the intensive care unit. *Computer Methods and Programs in Biomedicine* **209**, 106359 (2021). ISSN 0169-2607. doi: 10.1016/j.cmpb.2021.106359. URL <https://www.sciencedirect.com/science/article/pii/S0169260721004338>. Cited on page/s 7.
- [27] Andrea C. Skelly, Joseph R. Dettori, and Erika D. Brodt. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal* **3** (1), 9–12 (February 2012). ISSN 1663-7976 1869-4136. doi: 10.1055/s-0031-1298595. URL <https://pubmed.ncbi.nlm.nih.gov/23236300/>. Place: Germany. Cited on page/s 7.
- [28] Benjamin A. Goldstein, Nrupen A. Bhavsar, Matthew Phelan, and Michael J. Pencina. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *American journal of epidemiology* **184** (11), 847–855 (December 2016). ISSN 1476-6256 0002-9262. doi: 10.1093/aje/kww112. Place: United States. Cited on page/s 7.
- [29] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science* **356** (6334), 183–186 (2017). doi: 10.1126/science.aal4230. URL <https://www.science.org/doi/abs/10.1126/science.aal4230>. Cited on page/s 7.
- [30] Daniel Schönberger. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* **27** (2), 171–203 (May 2019). ISSN 0967-0769. doi: 10.1093/ijlit/eaz004. URL <https://doi.org/10.1093/ijlit/eaz004>. Cited on page/s 7.
- [31] Hossein Estiri, Zachary H. Strasser, Jeffery G. Klann, Thomas H. McCoy, Kavishwar B. Wagholikar, Sebastien Vasey, Victor M. Castro, MaryKate E. Murphy, and Shawn N. Murphy. Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations. *Patterns* **1** (4), 100051 (July 2020). ISSN 2666-3899. doi: 10.1016/j.patter.2020.100051. URL <https://www.sciencedirect.com/science/article/pii/S2666389920300623>. Cited on page/s 7.
- [32] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (April 2018). doi: 10.1136/bmj.k1479. URL <http://www.bmj.com/content/361/bmj.k1479.abstract>. Cited on page/s 8.

- [33] Ida Sim, Paul Gorman, Robert A. Greenes, R. Brian Haynes, Bonnie Kaplan, Harold Lehmann, and Paul C. Tang. Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *Journal of the American Medical Informatics Association* **8** (6), 527–534 (November 2001). ISSN 1067-5027. doi: 10.1136/jamia.2001.0080527. URL <https://doi.org/10.1136/jamia.2001.0080527>. Cited on page/s 8, 9.
- [34] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine* **3** (1), 17 (February 2020). ISSN 2398-6352. doi: 10.1038/s41746-020-0221-y. URL <https://doi.org/10.1038/s41746-020-0221-y>. Cited on page/s 8, 9.
- [35] Robert A. Greenes, David W. Bates, Kensaku Kawamoto, Blackford Middleton, Jerome Osherooff, and Yuval Shahar. Clinical Decision Support Models and Frameworks: Seeking To Address Research Issues Underlying Implementation Successes and Failures. *Journal of Biomedical Informatics* **78**, 134–143 (2018). doi: 10.1016/j.jbi.2017.12.005. URL <https://doi.org/10.1016/j.jbi.2017.12.005>. Cited on page/s 8.
- [36] Stijn Van de Velde, *et al.* A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. *Implementation Science* **13** (1) (2018). doi: 10.1186/s13012-018-0790-1. URL <https://doi.org/10.1186/s13012-018-0790-1>. Cited on page/s 8.
- [37] Rahul C. Deo. Machine Learning in Medicine. *Circulation* **132** (20), 1920–1930 (November 2015). ISSN 1524-4539 0009-7322. doi: 10.1161/CIRCULATIONAHA.115.001593. URL <https://pubmed.ncbi.nlm.nih.gov/26572668/>. Place: United States. Cited on page/s 8, 9.
- [38] Tiffani J. Bright, *et al.* Effect of clinical decision-support systems: a systematic review. *Annals of Internal Medicine* **157** (1), 29–43 (2012). doi: 10.7326/0003-4819-157-1-201207030-00450. URL <https://doi.org/10.7326/0003-4819-157-1-201207030-00450>. Cited on page/s 8.
- [39] David W. Bates, Gilad J. Kuperman, Samuel Wang, Tejal Gandhi, Anne Kittler, Lynn Volk, Cynthia Spurr, Ramin Khorasani, Milenko Tanasijevic, and Blackford Middleton. Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-Based Medicine a Reality. *Journal of the American Medical Informatics Association* **10** (6), 523–530 (2003). doi: 10.1197/jamia.m1370. URL <https://doi.org/10.1197/jamia.m1370>. Cited on page/s 8.
- [40] P. S. Roshanov, *et al.* Features of Effective Computerised Clinical Decision Support Systems: Meta-Regression of 162 Randomised Trials. *BMJ* **346** (feb14 1), f657–f657 (2013). doi: 10.1136/bmj.f657. URL <https://doi.org/10.1136/bmj.f657>. Cited on page/s 8.
- [41] Jan Horsky, Gordon D. Schiff, Douglas Johnston, Lauren Mercincavage, Douglas Bell, and Blackford Middleton. Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions. *Journal of Biomedical Informatics* **45** (6), 1202–1216 (2012). ISSN 1532-0464. doi: 10.1016/j.jbi.2012.09.002. URL <https://www.sciencedirect.com/science/article/pii/S1532046412001499>. Cited on page/s 8.
- [42] Mohamed Khalifa and Ibrahim Zabani. Improving Utilization of Clinical Decision Support Systems by Reducing Alert Fatigue: Strategies and Recommendations. *Studies in health technology and informatics* **226**, 51–54 (2016). ISSN 1879-8365 0926-9630. Place: Netherlands. Cited on page/s 9.
- [43] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association : JAMIA* **19** (1), 121–127 (February 2012). ISSN 1527-974X 1067-5027. doi: 10.1136/amiajnl-2011-000089. Place: England. Cited on page/s 9.
- [44] Elizabeth Murray, Joanne Burns, Carl May, Tracy Finch, Catherine O'Donnell, Paul Wallace, and Frances Mair. Why is it difficult to implement e-health initiatives? A qualitative

- study. *Implementation science : IS* **6**, 6 (January 2011). ISSN 1748-5908. doi: 10.1186/1748-5908-6-6. Place: England. Cited on page/s 9.
- [45] Joan S. Ash, Dean F. Sittig, Emily M. Campbell, Kenneth P. Guappone, and Richard H. Dykstra. Some unintended consequences of clinical decision support systems. *AMIA Annual Symposium Proceedings* **2007**, 26–30 (October 2007). ISSN 1942-597X 1559-4076. Place: United States. Cited on page/s 9.
- [46] Dawn Dowding, Natasha Mitchell, Rebecca Randell, Rebecca Foster, Valerie Lattimer, and Carl Thompson. Nurses’ use of computerised clinical decision support systems: a case site analysis. *Journal of clinical nursing* **18** (8), 1159–1167 (April 2009). ISSN 1365-2702 0962-1067. doi: 10.1111/j.1365-2702.2008.02607.x. Place: England. Cited on page/s 9.
- [47] Pranav Rajpurkar and Matthew P. Lungren. The Current and Future State of AI Interpretation of Medical Images. *New England Journal of Medicine* **388** (21), 1981–1990 (2023). doi: 10.1056/NEJMra2301725. URL <https://doi.org/10.1056/NEJMra2301725>. Cited on page/s 9.
- [48] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Working Paper 31422 National Bureau of Economic Research (July 2023). URL <http://www.nber.org/papers/w31422>. Series: Working Paper Series. Cited on page/s 9.
- [49] Ayush Jain, *et al.* Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. *JAMA Network Open* **4** (4), e217249–e217249 (April 2021). ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.7249. URL <https://doi.org/10.1001/jamanetworkopen.2021.7249>. Cited on page/s 9.
- [50] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* **25** (1), 30–36 (January 2019). ISSN 1546-170X. doi: 10.1038/s41591-018-0307-0. URL <https://doi.org/10.1038/s41591-018-0307-0>. Cited on page/s 9.
- [51] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”. *AI Magazine* **38** (3), 50–57 (September 2017). ISSN 0738-4602. doi: 10.1609/aimag.v38i3.2741. URL <https://doi.org/10.1609/aimag.v38i3.2741>. Publisher: John Wiley & Sons, Ltd. Cited on page/s 9.
- [52] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* **7** (2), 76–99 (May 2017). ISSN 2044-3994. doi: 10.1093/idpl/ix005. URL <https://doi.org/10.1093/idpl/ix005>. Cited on page/s 9.
- [53] Oskar Wysocki, Jessica Katharine Davies, Markel Vigo, Anne Caroline Armstrong, Dónal Landers, Rebecca Lee, and André Freitas. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence* **316**, 103839 (March 2023). ISSN 00043702. doi: 10.1016/j.artint.2022.103839. URL <https://linkinghub.elsevier.com/retrieve/pii/S0004370222001795>. Cited on page/s 9.
- [54] Anne F. Parkhill, Ornella Clavisi, Loyal Pattuwage, Marisa Chau, Tari Turner, Peter Bragge, and Russell Gruen. Searches for evidence mapping: effective, shorter, cheaper. *Journal of the Medical Library Association : JMLA* **99** (2), 157–160 (April 2011). ISSN 1558-9439 1536-5050. doi: 10.3163/1536-5050.99.2.008. Place: United States. Cited on page/s 10.
- [55] Sarah E. Hetrick, Alexandra G. Parker, Patrick Callahan, and Rosemary Purcell. Evidence mapping: illustrating an emerging methodology to improve evidence-based practice in youth mental health. *Journal of evaluation in clinical practice* **16** (6), 1025–1030 (December 2010). ISSN 1365-2753 1356-1294. doi: 10.1111/j.1365-2753.2008.01112.x. Place: England. Cited on page/s 10.

- [56] Peter Densen. Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association* **122**, 48–58 (2011). ISSN 0065-7778. Place: United States. Cited on page/s **10**.
- [57] Paul P Glasziou, Sharon Sanders, and Tammy Hoffmann. Waste in covid-19 research. *BMJ* **369**, m1847 (May 2020). doi: 10.1136/bmj.m1847. URL <http://www.bmj.com/content/369/bmj.m1847.abstract>. Cited on page/s **10**.
- [58] Perrine Créquit, Isabelle Boutron, Joerg Meerpohl, Hywel C. Williams, Jonathan Craig, and Philippe Ravaut. Future of evidence ecosystem series: 2. current opportunities and need for better tools and methods. *Journal of Clinical Epidemiology* **123**, 143–152 (July 2020). ISSN 0895-4356. doi: 10.1016/j.jclinepi.2020.01.023. URL <https://www.sciencedirect.com/science/article/pii/S0895435619305931>. Cited on page/s **10**.
- [59] James Thomas, *et al.* Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology* **91**, 31–37 (November 2017). ISSN 0895-4356. doi: 10.1016/j.jclinepi.2017.08.011. URL <https://www.sciencedirect.com/science/article/pii/S0895435617306042>. Cited on page/s **10**.
- [60] Kat Kolaski, Lynne Romeiser Logan, and John P.A. Ioannidis. Improving systematic reviews: guidance on guidance and other options and challenges. *Journal of Clinical Epidemiology* **159**, 266–273 (July 2023). ISSN 0895-4356. doi: 10.1016/j.jclinepi.2023.05.008. URL <https://www.sciencedirect.com/science/article/pii/S089543562300121X>. Cited on page/s **10**.
- [61] Carole Lunny, Cynthia Ramasubbu, Lorri Puil, Tracy Liu, Savannah Gerrish, Douglas M. Salzwedel, Barbara Mintzes, and James M. Wright. Over half of clinical practice guidelines use non-systematic methods to inform recommendations: A methods study. *PLOS ONE* **16** (4), e0250356 (April 2021). doi: 10.1371/journal.pone.0250356. URL <https://doi.org/10.1371/journal.pone.0250356>. Publisher: Public Library of Science. Cited on page/s **10**.
- [62] Hanan Khalil, Daniel Ameen, and Armita Zarnegar. Tools to support the automation of systematic reviews: a scoping review. *Journal of Clinical Epidemiology* **144**, 22–42 (April 2022). ISSN 0895-4356. doi: 10.1016/j.jclinepi.2021.12.005. URL <https://www.sciencedirect.com/science/article/pii/S0895435621004029>. Cited on page/s **10**.
- [63] David L. Sackett. Evidence-based medicine. *Seminars in Perinatology* **21** (1), 3–5 (February 1997). ISSN 0146-0005. doi: 10.1016/S0146-0005(97)80013-4. URL <https://www.sciencedirect.com/science/article/pii/S0146000597800134>. Cited on page/s **11**.
- [64] Steven H. Woolf, Richard Grol, Allen Hutchinson, Martin Eccles, and Jeremy Grimshaw. Potential benefits, limitations, and harms of clinical guidelines. *BMJ* **318** (7182), 527–530 (February 1999). ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.318.7182.527. URL <https://www.bmj.com/content/318/7182/527>. Publisher: British Medical Journal Publishing Group Section: Education and debate. Cited on page/s **11**.
- [65] Dean F. Sittig, Adam Wright, Jerome A. Osheroff, Blackford Middleton, Jonathan M. Teich, Joan S. Ash, Emily Campbell, and David W. Bates. Grand challenges in clinical decision support. *Journal of Biomedical Informatics* **41** (2), 387–392 (April 2008). ISSN 1532-0464. doi: 10.1016/j.jbi.2007.09.003. URL <https://www.sciencedirect.com/science/article/pii/S1532046407001049>. Cited on page/s **11**.
- [66] Thomas Brox Røst, Inger Dybdahl Sørby, and Gry Seland. Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-initiated Project. In *European Workshop on Practical Aspects of Health Informatics 2014* pages 53–62. Citeseer (2014). Cited on page/s **13**.
- [67] Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. Identifying catheter-related events through sentence classification. *International Journal of Data Mining and Bioinformatics* **23** (3), 213–233 (2020). doi: 10.1504/IJDMB.2020.107877. URL <https://dl.acm.org/doi/abs/10.1504/ijdmb.2020.107877>. Publisher: Inderscience Publishers (IEL). Cited on page/s **13**.

- [68] Thomas B. Røst, Laura Slaughter, Øystein Nytrø, Ashley E. Muller, and Gunn E. Vist. Using neural networks to support high-quality evidence mapping. *BMC Bioinformatics* **22** (11), 496 (October 2021). ISSN 1471-2105. doi: 10.1186/s12859-021-04396-x. URL <https://doi.org/10.1186/s12859-021-04396-x>. Cited on page/s 13.
- [69] Thomas Brox Røst, Carolyn Clausen, Øystein Nytrø, Roman Kuposov, Bennett Leventhal, Odd Sverre Westbye, Victoria Bakken, Linda Helen Knudsen Flygel, Kaban Koochakpour, and Norbert Skokauskas. Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services. *Frontiers in Psychiatry* **11** (2020). ISSN 1664-0640. doi: 10.3389/fpsy.2020.564205. URL <https://www.frontiersin.org/articles/10.3389/fpsy.2020.564205>. Cited on page/s 13.
- [70] Carolyn Clausen, Bennett Leventhal, Øystein Nytrø, Roman Kuposov, Thomas Brox Røst, Odd Sverre Westbye, Kaban Koochakpour, Thomas Frodl, Line Stien, and Norbert Skokauskas. Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development. *Frontiers in Psychiatry* **14** (2023). ISSN 1664-0640. doi: 10.3389/fpsy.2023.1033724. URL <https://www.frontiersin.org/articles/10.3389/fpsy.2023.1033724>. Cited on page/s 13.
- [71] EVICARE - Evidence-based care processes: Integrating knowledge in clinical information systems (2009). URL <https://prosjektbanken.forskningsradet.no/project/FORISS/193022>. Cited on page/s 14.
- [72] BIGMED: A big data medical solution for precision medicine. - Prosjektbanken (2016). URL <https://prosjektbanken.forskningsradet.no/project/FORISS/259055>. Cited on page/s 14.
- [73] Individualized Digital DEcision Assist System (IDDEAS) for the diagnosis and management of mental and behavior disorders in children and adolescents (2017). URL <https://prosjektbanken.forskningsradet.no/project/FORISS/269117>. Cited on page/s 14.
- [74] Thomas Brox Røst, Ola Huseth, Øystein Nytrø, and Anders Grimsmo. Lessons from Developing an Annotated Corpus of Patient Histories. *J. Comput. Sci. Eng.* **2** (2), 162–179 (2008). Cited on page/s 14.
- [75] Thomas Brox Røst, Ole Edsberg, Anders Grimsmo, and Øystein Nytrø. Comparing medical code usage with the compression-based dissimilarity measure. *Studies in Health Technology and Informatics* **129** (Pt 1), 684–688 (2007). ISSN 0926-9630. Cited on page/s 14.
- [76] Thomas Brox Røst, Øystein Nytrø, and Anders Grimsmo. Classifying encounter notes in the primary care patient record. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval* volume 205 pages 1–5 (2006). Cited on page/s 15.

CHAPTER 2

Results

This section summarizes the papers included in the thesis and their contributions towards the research questions.

RQ1 How can health-related data and knowledge aid decision-making?

RQ2 How can healthcare professionals contribute to implementation of data-driven decision-making?

RQ3 How can patient trajectory data contribute to decision-making in healthcare?

2.1 PAPER A: DEVELOPMENT OF A MEDICATION RECONCILIATION TOOL FOR NORWEGIAN PRIMARY CARE EPR SYSTEMS: EXPERIENCES FROM A USER-INITIATED PROJECT

Authors: Thomas Brox Røst, Inger Dybdahl Sørby, Gry Seland

Published in: *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*. Trondheim, Norway, May 19th 2014.¹

Research Questions: RQ1, RQ2

Abstract: Medication reconciliation is one of the most important priorities of national and international patient safety efforts, due to the numerous deaths and adverse drug reactions caused by inappropriate medication use. One of the main challenges of general practitioners (GPs) is to get an overview of changes in the patients' medications after transitions between healthcare institutions. This paper presents how Natural Language Processing of free text notes such as discharge summaries is used to automatically extract information about medications and how this can be compared to the patient's existing medication list in an electronic patient record (EPR) system. The functionality has been developed in a user initiated project, as a cooperation between four different vendors and with a strong involvement of the end users. The functionality

Legemidler i bruk						Medikament fra tekst		
Vurdert	Kategori	Navn	Virkestoff	Styrke	Dosering	Navn	Styrke	Dosering
<input checked="" type="checkbox"/>	Fast	Albyl-E	Acetylsalisyls...	75mg	1 tbl dgl	Albyl-E	75mg	x 1
<input checked="" type="checkbox"/>	Fast	Orfiril	Valproinsyre	300mg	1 tbl 1 ggr dgl	Orfiril	300mg	x 1
<input checked="" type="checkbox"/>						Cipramil	10mg	10 mg
<input checked="" type="checkbox"/>						Zopiklon Mylan	5mg	
<input checked="" type="checkbox"/>	Kur	Apocilin	Fenoksymetyl...	660mg	1 + 1 + 2	Apocilin	660mg	1 + 1 + 2
<input type="checkbox"/>						Lamotrigin	25mg	x 2
Inngående tekst til ViVit fra Infodoc Plenario						Tekst etter tolkning av ViVit		
Cipramil Tab 10 mg 10mg Tablett - Blisterpakning Dsnn 1 tbl dgl (Fast)						Cipramil Tab 10 mg 10 mg Tablett - Blisterpakning Dsnn 1 tbl dgl (Fast)		
Zopiklon Mylan 5mg Tablett - Blisterpakning Dsnn 1 tbl kveld (Behov)						Zopiklon Mylan 5 mg Tablett - Blisterpakning Dsnn 1 tbl kveld (Behov)		
Apocilin 660mg 1+1+2 (kur)						Apocilin 660 mg 1 + 1 + 2 (kur)		
Albyl-E 75mgx1						Albyl-E 75 mg x 1		
Orfiril tbl 300mgx1 (NY)						Orfiril tbl 300 mg x 1 (NY)		
Lamotrigin tbl 25mg x 2 (NY)						Lamotrigin tbl 25 mg x 2 (NY)		

FIGURE 2.1. Screenshot of the medication reconciliation tool in use within the Infodoc EPR system.

is available for most Norwegian GPs and is seen as a very useful tool in the medication reconciliation process.

2.1.1 Paper Summary

This study describes the process and trade-offs required for planning, implementing and deploying an NLP-based tool for medication reconciliation within the Norwegian healthcare system. Medication reconciliation²⁻⁴ is the process where a patient's general practitioner correlates the information about medications and prescriptions found in the hospital discharge note with the information in the general practice (GP) EPR system. At the time of implementation, the discharge note was transmitted electronically but not in a structured format. For this reason, the medication reconciliation process was seen as time-consuming and prone to errors, but also as a necessary administrative task for keeping the record systems in sync. We developed a system for parsing discharge note text for mentions of medications along with dosage and frequency, mapping the found medications to the Norwegian FEST medication database and ATC codes, and attempting to match the medications with those already existing in the general practice EPR. The system was eventually deployed as a standalone module within the three major Norwegian GP EPR systems and is still in use at the time of writing.

2.2 PAPER B: IDENTIFYING CATHETER-RELATED EVENTS THROUGH SENTENCE CLASSIFICATION

Authors: Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, Øystein Nytrø

Published in: *International Journal of Data Mining and Bioinformatics* 23 (3), 213-233 (2020). Inderscience Publishers (IEL.)⁵

Research Questions: RQ1, RQ2, RQ3

Abstract: Infections caused by central venous catheter (CVC) use is a serious and under-reported problem in healthcare. The CVC is almost ubiquitous in critical care because it enables fast circulatory monitoring and central administration of medication and nutrition. However, the CVC exposes the patient to a risk of blood-stream infections (BSI). Explicit documentation of normal CVC usage and exposure is sparse and indirect in the health record. For a clinician, CVC presence is simple to infer from record statements about procedures, plans and results related to CVC. In order to capture evidence about CVC-related risk of infections and complications, it is important to develop computerized tools that can estimate individual patient days of CVC exposure retrospectively for large cohorts of patients. Towards that objective, we have developed methods for learning classifiers for statements about CVC-related events occurring in the textual health record. This includes developing and testing an annotation ontology of events and indicators, annotation guidelines, a gold standard of annotated clinical records selected from a corpus of complete health records for more 800 episodes of care and collecting alternate health register evidence for validation purposes. This paper describes the available data and gold standard, feature selection approaches and our experiments with different classification algorithms. We find that even with limited data it is possible to build reasonably accurate sentence classifiers for the most important events. We also find that making use of document meta information helps improve classification quality by providing additional context to a sentence. Finally, we outline some strategies on using our results for future analysis and reasoning about CVC usage intervals and CVC exposure over individual patient trajectories.

2.2.1 Paper Summary

Blood-stream infection as a result of central venous catheter (CVC) use is a serious problem in healthcare, with a high risk of complications and in worst case death⁶. It is also known to be an under-reported problem⁷ and manual surveillance regimes are costly and difficult to implement properly. This study wanted to investigate if automated detection of CVC-related events from clinical notes in the patient record was possible. A corpus consisting of more than 800 patient episodes of care was collected and annotated with information about events connected with the use of CVC in patient treatment. We found that even with a relatively small data set it was possible to build machine learning classifiers that detected key events with reasonable precision and recall.

2.3 PAPER C: USING NEURAL NETWORKS TO SUPPORT HIGH-QUALITY EVIDENCE MAPPING

Authors: Thomas Brox Røst, Laura Slaughter, Øystein Nytrø, Ashley Elizabeth Muller, Gunn Elisabeth Vist

Published in: *BMC Bioinformatics* 22 (11), 496 (October 2021).⁸

Research Questions: RQ1, RQ2

Abstract:

Background: The Living Evidence Map Project at the Norwegian Institute of Public Health (NIPH) gives an updated overview of research results and publications. As part of NIPH's mandate to inform evidence-based infection prevention, control and treatment, a large group of experts are continuously monitoring, assessing, coding and summarising new COVID-19 publications. Screening tools, coding practice and workflow are incrementally improved, but remain largely manual.

Results: This paper describes how deep learning methods have been employed to learn classification and coding from the steadily growing NIPH COVID-19 dashboard data, so as to aid manual classification, screening and preprocessing of the rapidly growing influx of new papers on the subject. Our main objective is to make manual screening scalable through semi-automation, while ensuring high-quality Evidence Map content.

Conclusions: We report early results on classifying publication topic

and type from titles and abstracts, showing that even simple neural network architectures and text representations can yield acceptable performance.

2.3.1 Paper Summary

This study was motivated by a collaboration with the Living Evidence Map Project⁹ at the Norwegian Institute of Public Health (NIPH), which in turn came about as an attempt to keep up with the large volume of COVID-19-related scientific publications following the March 2020 declaration of a global pandemic. Researchers and staff at NIPH were using mostly manual methods to screen and classify new research and found it hard to keep up with the inflow of publications. For this reason they wanted to explore if the use of machine learning could be used to reduce the effort and resources needed when producing systematic reviews and evidence maps while still maintaining quality and precision.

Using a manually created data set based on a coding manual produced by NIPH we explored various ways of classifying paper topics and types based on their titles and abstracts. The conclusion was that even simple deep learning approaches using quite sparse data could produce results worth pursuing further.

2.4 PAPER D: LOCAL, EARLY, AND PRECISE: DESIGNING A CLINICAL DECISION SUPPORT SYSTEM FOR CHILD AND ADOLESCENT MENTAL HEALTH

Authors: Thomas Brox Røst, Carolyn Clausen, Øystein Nytrø, Roman Koposov, Bennett Leventhal, Odd Sverre Westbye, Victoria Bakken, Linda Helen Knudsen Flygel, Kaban Koochakpour, Norbert Skokauskas

Published in: *Frontiers in Psychiatry* 11 (2020).¹⁰

Research Questions: RQ1, RQ3

Abstract: Mental health disorders often develop during childhood and adolescence, causing long term and debilitating impacts at individual and societal levels. Local, early, and precise assessment and evidence-based treatment are key to achieve positive mental health outcomes and to avoid long-term care. Technological advancements, such as computerized Clinical Decision Support Systems (CDSSs), can support practitioners in providing evidence-based care. While previous studies have found CDSS implementation

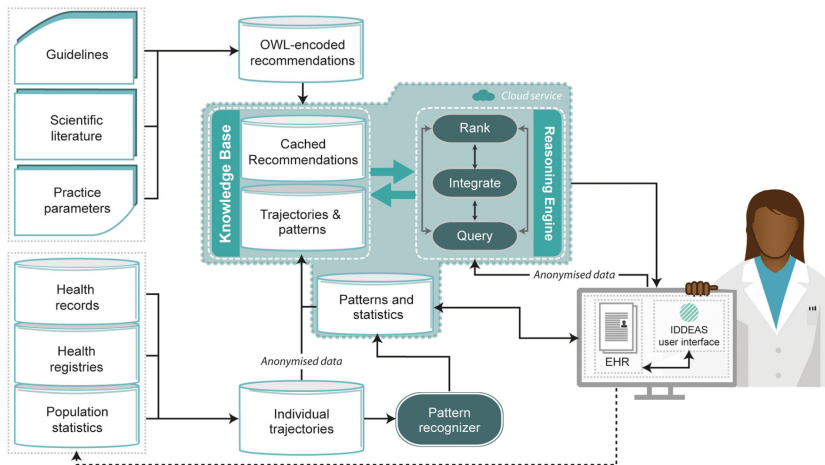


FIGURE 2.2. The IDDEAS clinical decision support model.

helps to improve aspects of medical care, evidence is limited on its use for child and adolescent mental health care. This paper presents challenges and opportunities for adapting CDSS design and implementation to child and adolescent mental health services (CAMHS). To highlight the complexity of incorporating CDSSs within local CAMHS, we have structured the paper around four components to consider before designing and implementing the CDSS: supporting collaboration among multiple stakeholders involved in care; optimally using health data; accounting for comorbidities; and addressing the temporality of patient care. The proposed perspective is presented within the context of the child and adolescent mental health services in Norway and an ongoing Norwegian innovative research project, the Individualized Digital DEcision Assist System (IDDEAS), for child and adolescent mental health disorders. Attention deficit hyperactivity disorder (ADHD) among children and adolescents serves as the case example. The integration of IDDEAS in Norway intends to yield significantly improved outcomes for children and adolescents with enduring mental health disorders, and ultimately serve as an educational opportunity for future international approaches to such CDSS design and implementation.

2.4.1 Paper Summary

This paper was supported by the Individualized Digital DEcision Assist System (IDDEAS) project¹¹, where the aim was to design and implement clinical

decision-support systems (CDSS) for supporting diagnosis and treatment of ADHD in child and adolescent mental health services (CAMHS). CDSS in medicine is a well-explored concept but not so in CAMHS¹². In addition, the recent rise in mental health disorders among children and adolescents makes this a worthy area of study^{9,13}. In the paper we identified some key considerations for successful implementation of CDSS in CAMHS: support for collaboration among multiple stakeholders, data-driven decision-making, taking comorbidities into account, and ensuring that temporality of care is addressed.

2.5 PAPER E: USABILITY OF THE IDDEAS PROTOTYPE IN CHILD AND ADOLESCENT MENTAL HEALTH SERVICES: A QUALITATIVE STUDY FOR CLINICAL DECISION SUPPORT SYSTEM DEVELOPMENT

Authors: Carolyn Clausen, Bennett Leventhal, Øystein Nytrø, Roman Koposov, Thomas Brox Røst, Odd Sverre Westbye, Kaban Koochakpour, Thomas Frodl, Line Stien, Norbert Skokauskas

Published in: *Frontiers in Psychiatry* 14 (2023).¹⁴

Research Questions: RQ1, RQ2

Abstract:

Introduction: Child and adolescent mental health services (CAMHS) clinical decision support system (CDSS) provides clinicians with real-time support as they assess and treat patients. CDSS can integrate diverse clinical data for identifying child and adolescent mental health needs earlier and more comprehensively. Individualized Digital Decision Assist System (IDDEAS) has the potential to improve quality of care with enhanced efficiency and effectiveness.

Methods: We examined IDDEAS usability and functionality in a prototype for attention deficit hyperactivity disorder (ADHD), using a user-centered design process and qualitative methods with child and adolescent psychiatrists and clinical psychologists. Participants were recruited from Norwegian CAMHS and were randomly assigned patient case vignettes for clinical evaluation, with and without IDDEAS. Semi-structured interviews were conducted as one part of testing the usability of the prototype following a five-question interview guide. All interviews were recorded, transcribed, and analyzed following qualitative content analysis.

Tilbake ✕

BAKGRUNNSINFORMASJON

Bruk av retningslinje for ADHD

Anbefalingene i dette systemet er basert på standardiserte kliniske retningslinjer (ICD-10) for å støtte deg i å ta beslutninger. Velg "ja" eller "nei" avhengig av den tilgjengelige pasientinformasjonen. Hvis du ikke har tilstrekkelig informasjon for å gi et svar så bruker du "nei".

FORTSETT

Tilbake ✕

Utredningskriterier

Følgende utredningskriterier må brukes.

Symptomene startet før 12 års alderen. Nei Ja

Symptomene påvirker daglig fungering. Nei Ja

Tilstanden er tilstede i to eller flere settinger. Nei Ja

FORTSETT

Tilbake ✕


Avgjør om vurderingskriteriene er fulgt

Gitt de valgene du har gjort for denne pasienten og de standardiserte kliniske retningslinjene i IDDEAS er dette den beste anbefalingen:

Utredningen er fullført.

Du kan enten godta beslutningsanbefalingen for å fortsette retningslinjen eller velge "Avbryt" hvis du er uenig.

AVBRYT **GODTA ANBEFALING**

 **Pasient D**
Alder: 10

10 år gammel gutt D har hatt problemer med å høre etter siden han startet på skolen. Han er impulsiv og distraheres lett. Han beveger seg konstant i timene, bråker og bryter regler. Han presterer dårlig på skolen. Han prøver å få venner, men de andre barna har ikke lyst til å leke med han fordi de synes han er «rar». Han prøver hardt på skolen, men greier ikke oppgavene sine. Foreldrene rapporterer om lignende problemer på hjemmebane. Han fullfører ikke daglige rutiner selv med mange påminnelser. Foreldrene hans prøver å hjelpe han med leksene, men han sliter med oppgavene. Han er ofte i slåsskamp med nabobarna, og må dermed for det meste holde seg hjemme. Han er kjempegod i videospill og spiller ofte med online venner.

FIGURE 2.3. Individualized Digital Decision Assist System prototype software screenshot.

Results: Participants were the first 20 individuals from the larger IDDEAS prototype usability study. Seven participants explicitly stated a need for integration with the patient electronic health record system. Three participants commended the step-by-step guidance as potentially helpful for novice clinicians. One participant did not like the aesthetics of the IDDEAS at this stage. All participants were pleased about the display of the patient information along with guidelines and suggested that wider guideline coverage will make IDDEAS much more useful. Overall, participants emphasized the importance of maintaining the clinician as the decision-maker in the clinical process, and the overall potential utility of IDDEAS within Norwegian CAMHS.

Conclusion: Child and adolescent mental health services psychiatrists and psychologists expressed strong support for the IDDEAS clinical decision support system if better integrated in daily workflow. Further usability assessments and identification of additional IDDEAS requirements are necessary. A fully functioning, integrated version of IDDEAS has the potential to be an important support for clinicians in the early identification of risks for youth mental disorders and contribute to improved assessment and treatment of children and adolescents.

2.5.1 Paper Summary

In this second paper supported by the IDDEAS project we tested a prototype CDSS for CAMHS on a selection of child and adolescent psychiatrists and clinical psychologists. After being randomly assigned a patient case vignette for clinical evaluation within IDDEAS, the participants underwent a semi-structured interview about their experiences regarding the usability of the prototype. Feedback included the need for integration with electronic health record systems, the usefulness of step-by-step guidance and the combination of patient information with guidelines, the need for additional guideline coverage, and the importance of maintaining the clinician as the decision-maker in the clinical process. Overall, participants expressed support for CDSS for CAMHS as a concept and the potential of the IDDEAS CDSS as part of a future workflow for early identification and improved treatment of youth mental disorders.

2.6 REFERENCES

- [1] Thomas Brox Røst, Inger Dybdahl Sørby, and Gry Seland. Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-initiated Project. In *European Workshop on Practical Aspects of Health Informatics 2014* pages 53–62. Citeseer (2014). Cited on page/s 23.

- [2] Peter Pronovost, Brad Weast, Mandalyn Schwarz, Rhonda M Wyskiel, Donna Prow, Shelley N Milanovich, Sean Berenholtz, Todd Dorman, and Pamela Lipsett. Medication reconciliation: a practical tool to reduce the risk of medication errors. *Journal of Critical Care* **18** (4), 201–205 (December 2003). ISSN 0883-9441. doi: 10.1016/j.jcrc.2003.10.001. URL <https://www.sciencedirect.com/science/article/pii/S0883944103001084>. Cited on page/s 24.
- [3] Stephanie K. Mueller, Kelly Cunningham Sponsler, Sunil Kripalani, and Jeffrey L. Schnipper. Hospital-Based Medication Reconciliation Practices: A Systematic Review. *Archives of Internal Medicine* **172** (14), 1057–1069 (July 2012). ISSN 0003-9926. doi: 10.1001/archinternmed.2012.2246. URL <https://doi.org/10.1001/archinternmed.2012.2246>. Cited on page/s 24.
- [4] Elin C. Lehnbohm, Michael J. Stewart, Elizabeth Manias, and Johanna I. Westbrook. Impact of Medication Reconciliation and Review on Clinical Outcomes. *Annals of Pharmacotherapy* **48** (10), 1298–1312 (October 2014). ISSN 1060-0280. doi: 10.1177/1060028014543485. URL <https://doi.org/10.1177/1060028014543485>. Publisher: SAGE Publications Inc. Cited on page/s 24.
- [5] Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. Identifying catheter-related events through sentence classification. *International Journal of Data Mining and Bioinformatics* **23** (3), 213–233 (2020). doi: 10.1504/IJDMB.2020.107877. URL <https://dl.acm.org/doi/abs/10.1504/ijdmb.2020.107877>. Publisher: Inderscience Publishers (IEL). Cited on page/s 25.
- [6] Robert W. Taylor and Ashok V. Palagiri. Central Venous Catheterization. *Critical Care Medicine* **35** (5), 1390–1396 (2007). doi: 10.1097/01.ccm.0000260241.80346.1b. URL <https://doi.org/10.1097/01.ccm.0000260241.80346.1b>. Cited on page/s 26.
- [7] Adrian VK Wong, *et al.* Insertion rates and complications of central lines in the UK population: A pilot study. *Journal of the Intensive Care Society* **19** (1), 19–25 (2018). doi: 10.1177/1751143717722914. URL <https://doi.org/10.1177/1751143717722914>. Cited on page/s 26.
- [8] Thomas B. Røst, Laura Slaughter, Øystein Nytrø, Ashley E. Muller, and Gunn E. Vist. Using neural networks to support high-quality evidence mapping. *BMC Bioinformatics* **22** (11), 496 (October 2021). ISSN 1471-2105. doi: 10.1186/s12859-021-04396-x. URL <https://doi.org/10.1186/s12859-021-04396-x>. Cited on page/s 26.
- [9] Norwegian Institute of Public Health. A systematic and living evidence map on COVID-19 (2020). URL <https://www.fhi.no/contentassets/e64790be5d3b4c4abe1f1be25fc862ce/covid-19-evidence-map-protocol-20200403.pdf>. Cited on page/s 27, 29.
- [10] Thomas Brox Røst, Carolyn Clausen, Øystein Nytrø, Roman Koposov, Bennett Leventhal, Odd Sverre Westbye, Victoria Bakken, Linda Helen Knudsen Flygel, Kaban Koochakpour, and Norbert Skokauskas. Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services. *Frontiers in Psychiatry* **11** (2020). ISSN 1664-0640. doi: 10.3389/fpsy.2020.564205. URL <https://www.frontiersin.org/articles/10.3389/fpsy.2020.564205>. Cited on page/s 27.
- [11] Individualized Digital DEcision Assist System (IDDEAS) for the diagnosis and management of mental and behavior disorders in children and adolescents (2017). URL <https://prosjektbanken.forskningsradet.no/project/FORISS/269117>. Cited on page/s 28.
- [12] Roman Koposov, *et al.* Clinical Decision Support Systems in Child and Adolescent Psychiatry: a Systematic Review. *European Child & Adolescent Psychiatry* **26** (11), 1309–1317 (2017). doi: 10.1007/s00787-017-0992-0. URL <https://doi.org/10.1007/s00787-017-0992-0>. Cited on page/s 29.
- [13] Norbert Skokauskas, M. Diane Eckert, Gerald Busch, Joy K. L. Andrade, Taryn M. Park, and Anthony P. S. Guerrero. Sustainable child and adolescent psychiatry. *International Review*

- of Psychiatry* **34** (2), 97–100 (February 2022). ISSN 0954-0261. doi: 10.1080/09540261.2022.2082163. URL <https://doi.org/10.1080/09540261.2022.2082163>. Publisher: Taylor & Francis. Cited on page/s 29.
- [14] Carolyn Clausen, Bennett Leventhal, Øystein Nytrø, Roman Kuposov, Thomas Brox Røst, Odd Sverre Westbye, Kaban Koochakpour, Thomas Frodl, Line Stien, and Norbert Skokauskas. Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development. *Frontiers in Psychiatry* **14** (2023). ISSN 1664-0640. doi: 10.3389/fpsy.2023.1033724. URL <https://www.frontiersin.org/articles/10.3389/fpsy.2023.1033724>. Cited on page/s 29.

CHAPTER 3

Discussion

In this chapter, each publication and its contribution to the research questions is discussed. We also reflect on possible future work and make some broader observations.

RQ1 How can health-related data and knowledge aid decision-making?

RQ2 How can healthcare professionals contribute to implementation of data-driven decision-making?

RQ3 How can patient trajectory data contribute to decision-making in health-care?

3.1 PAPER A: DEVELOPMENT OF A MEDICATION RECONCILIATION TOOL FOR NORWEGIAN PRIMARY CARE EPR SYSTEMS: EXPERIENCES FROM A USER-INITIATED PROJECT

This paper gives insight into a process where medical language processing technology was integrated into a general practice workflow to reduce the time spent on medication reconciliation and to improve the quality of the output of this process (**RQ1**). The project was a product of national funding initiatives and ICT in medicine improvement programs (specifically "EPJ-løftet"¹) that were actively seeking smart, light-weight technology solutions to known problems and challenges in general practice. A key success factor was that the project was initiated by end users and stakeholders, in this case the Norwegian College of General Practice ("Norsk forening for allmennmedisin"), who had specific ideas about how technology could improve their daily work with patients (**RQ2**). The participation of GPs was not only important in them being correctly positioned to drive such initiatives forward but also in active participation in all parts of the project from planning, specification, data collection and testing. Throughout the lifetime of the project funding was provided from the Directorate of Health for continuously updating the gold standard training data set so that the tool could evolve with changing documentation practices.

Another issue encountered during the project planning and implementation phase were the technical and legal challenges involved with making the technology work in a healthcare setting with real patients. Since potentially sensitive patient data could be transmitted, both health registry and data privacy laws would apply. The only realistic solution was to bundle the technology as a standalone component that would run locally on the EPR installation in each GP office. Moreover, this component would need to interface with at least three different technology platforms, whose capabilities would differ and with different assumptions regarding the processing power available at each EPR installation site. The component would have to be fully self-sufficient with no transmission of error and operating data back to the developers. This ended up making the development, distribution and maintenance process much more complicated than in an ideal implementation situation—but at the same time highlighting the constraints, issues and challenges found when putting language processing technologies to use in a real-world healthcare setting.

3.2 PAPER B: IDENTIFYING CATHETER-RELATED EVENTS THROUGH SENTENCE CLASSIFICATION

A key motivation for this project was to use the detected CVC events to improve our knowledge about CVC use in hospitals. An important metric for this purpose would be the duration of CVC use, i.e., the number of CVC days. Linking this information with the prevalence of catheter-related blood-stream infections would provide useful information for targeted quality improvement work (**RQ1**). Events related to use of CVC (which were also most common in the data set) had the best overall precision and recall prediction quality. While we did not infer patient CVC days in this study, the results indicate that this should be possible, given access to sufficiently complete patient histories. This also highlights another aspect of this study, which was the application of annotation and machine learning methods on longitudinal and heterogeneous patient records, i.e., different record note types spread out over a period covering admission, treatment and discharge. We found that both the use of structural information within the note and the note type (e.g. nursing notes, discharge notes or anesthesiology record notes) itself could in some cases be used to improve the classification quality. In ongoing unpublished work we have seen that considering the note type and its position in the patient history is indeed often very useful information when trying to infer CVC use days from discovered CVC use events in the patient history. This speaks to the potential for using complete patient histories with heterogeneous clinical note types for medical information extraction tasks (**RQ3**). The participation of a nurse with special competence in infection control was another important factor.

CVC-related events were in many cases more implicit than explicit, requiring a trained clinician to be able to infer from the clinical text records what events had likely taken place (RQ2).

3.3 PAPER C: USING NEURAL NETWORKS TO SUPPORT HIGH-QUALITY EVIDENCE MAPPING

The motivation behind this publication was to investigate the possibility of introducing machine learning as part of a coding and review workflow (RQ1). The most interesting result was the knowledge gained about how an ongoing coding and systematization effort operates and how this should guide the eventual introduction of automation as part of the workflow (RQ2). The experiments were done during an ongoing project and not long after the fact, as is often common when creating gold standard data sets for medical informatics research. The coding manual was an evolving document and coding standards were likely to (and indeed did) change during the duration of the project. This yielded insights such as the need to have the machine learning models evolve with the workflow. Moreover, manual coding is, even with a coding manual, a process where disagreement is normal. In the NIPH coding workflow, disagreements were usually resolved but this does not always imply that the agreed-upon resolution is more correct than the alternatives. As with any decision support application, automation must fit into a workflow that supports manual review, disagreement resolution, and oversight.

3.4 PAPER D: LOCAL, EARLY, AND PRECISE: DESIGNING A CLINICAL DECISION SUPPORT SYSTEM FOR CHILD AND ADOLESCENT MENTAL HEALTH

The main contribution of this paper was in contrasting the traditional application of CDSS in healthcare with the practical considerations when providing care to CAMHS patients. Treatment of children and adolescents with mental health problems will typically involve collaboration between multiple stakeholders, such as parents, schools and counseling services. These will for good reasons have different perspectives on the patient and have responsibility for different aspects of the treatment. Their information models will differ and can provide different but equally valuable background to patient treatment (RQ1). Moreover, comorbidities are common in the targeted patient group and a CDSS must take into account that a narrow focus on support for a specific diagnosis may not be feasible. Finally, the longitudinal aspect of CAMHS treatment must be catered for (RQ3). Taken together, these considerations expand on the

known challenges of making CDSS work in actual clinical use, in particular for complex care situations where multiple stakeholders are involved.

3.5 PAPER E: USABILITY OF THE IDDEAS PROTOTYPE IN CHILD AND ADOLESCENT MENTAL HEALTH SERVICES: A QUALITATIVE STUDY FOR CLINICAL DECISION SUPPORT SYSTEM DEVELOPMENT

While the tested system was an early prototype, it still showed the usefulness of applying user-centered design principles to a CDSS for complex care situations. Given the problems often encountered when applying CDSS to clinical practice, the experiment provided important feedback in the form of learning about clinicians' preferences and their workflow in a diagnostic setting (RQ2). A key takeaway was the necessity of close integration with existing electronic patient records. Access to existing patient information is a critical usability factor, both from the user point of view in terms of looking up information needed for decision-making but also from a system point of view in the sense of making relevant information automatically available to the clinician. The study confirmed the need for patient information from other stakeholders (e.g. schools and psychological counseling services), as put forward in paper D (RQ1). As this study focused on a particular type of stakeholder, i.e., the specialist clinicians, focus was naturally on their needs and preferences. While they maintained their need for autonomy in the decision-making process, this does not necessarily preclude a workflow where other stakeholders are more visible. Early risk identification is known to be a key factor in CAMHS when it comes to prevention of mental health disorders². Furthermore, precise and early risk identification often depends on relevant information from school, parents and primary care. This highlights the need to think about the operating context of CDSS in a broader scope, utilizing information flow from multiple stakeholders to provide efficient and precise provisioning of early intervention.

3.6 REFERENCES

- [1] Direktoratet for e helse. EPJ-løftet (2023). URL <https://www.ehelse.no/programmer/epj-loftet>. Cited on page/s 35.
- [2] Filipa Sampaio, Inna Feldman, Tara A. Lavelle, and Norbert Skokauskas. The cost-effectiveness of treatments for attention deficit-hyperactivity disorder and autism spectrum disorder in children and adolescents: a systematic review. *European Child & Adolescent Psychiatry* 31 (11), 1655–1670 (November 2022). ISSN 1435-165X. doi: 10.1007/s00787-021-01748-z. URL <https://pubmed.ncbi.nlm.nih.gov/33751229/>. Cited on page/s 38.

CHAPTER 4

Conclusion

When reflecting on the changes that have taken place since the work in this thesis was started, the most striking has been the advances within machine learning (ML) and natural language processing (NLP). The recent excitement around *foundation models*¹ is the current culmination of a line of developments that started in the 1980s with the introduction of learning algorithms within the field of artificial intelligence (AI). It was now no longer necessary to specify exactly how to solve a problem but rather let the learning algorithm figure it out by itself. This meant that the same algorithms could tackle a range of different problems merely by changing the data the algorithms were trained on. However, feature engineering and domain knowledge was still necessary for complex tasks and models were usually supervised and highly specialized to the task at hand.

4.1 DEEP LEARNING AND FOUNDATION MODELS

The rise of deep learning from around 2010 led to a marked change in the capabilities and scope of machine learning models², starting with considerable performance gains in image classification tasks by Krizhevsky et al.³ and more recently from the introduction of transfer learning-driven foundation models¹. Foundation models, which include large language models (LLM) such as GPT-4, are self-supervised machine learning models trained on a broad range of data at large scale. Instead of being tailored to solving specific tasks, these models can perform better than previous generations of ML models on a number of different problems. This is achieved by a combination of large-scale computing and hardware innovations, access to vast volumes of training data, and the invention of the transformer model architecture by Vaswani et al.⁴ which opened up for better hardware utilization during model training.

Within healthcare the shift to foundation models has, apart from some early efforts, yet to take place according to Moor et al.⁵ Access to large-scale, diverse medical datasets continues to be an obstacle, as previously described. In addition, the complexity of the medical domain is another challenge. Medical

ML models are for now highly task-specific, inflexible, and mostly supervised. However, more generalist medical machine learning models with dynamic reasoning capabilities are likely to emerge.

4.2 INTERFACE EXPECTATIONS

An interesting aspect of the rise of foundational LLMs such as the GPT class of models⁶ is how the introduction of a user-friendly low-threshold interface in the form of ChatGPT was an enabler for a major shift in both public perception and use of advanced language models. Aside from the model's ability to pass medical exams^{7,8} and to be perceived as more empathetic than human doctors⁹, a generation of users are learning to query advanced medical models through chat, dialogue, and prompt engineering. For all the faults and limitations of the current generation of ML models, a new standard has been set in terms of how to interact with tools that help users explore knowledge and, ultimately, make decisions. This further raises the bar for decision-support applications in the medical domain, where user-friendliness and workflow process fit is a known success factor. It is possible that future users of CDSS will consider the availability of natural language dialogue-based interfaces as a fundamental requirement for many application types.

4.3 OBSTACLES TO AI IN MEDICINE

Lee et al. argue that current LLMs have several problems that limit their applicability for the healthcare domain¹⁰. For example, their ability to answer questions about clinical knowledge is still inferior to that of a clinician, as seen in the work of Singhal et al.¹¹ Given the safety-critical nature of patient treatment, models that may generate medical misinformation or reproduce harmful biases from the source data need to be used with care. It is possible to improve a model's ability to avoid generating harmful content, as has been shown with GPT-4¹², but so far there are several ethical concerns related to the use of such models¹³. While question-answering performance on standard patient vignettes is impressive¹⁴, correctly and consistently adapting answers to specific patient circumstances remains a challenge¹⁵. When targeting patients instead of clinicians, Nov et al. found that patients already find it difficult to differentiate answers given by ChatGPT from those created by clinicians¹⁶. This is remarkable in itself but further complicates the use of such models for e.g. decision-making and patient communication¹³. The previously mentioned concerns around the explainability of model-generated answers also remain¹. In any case, the step from research project and artificial settings to clinical

deployment for patient treatment is still a large one. In a review by Vornow et al., most existing medical LLMs are found to be evaluated on tasks that say little about their usefulness to healthcare systems¹⁷. The need for measuring the effect of LLM applications on factors such mortality and patient outcome in clinical interventions is as relevant as ever.

There are also barriers to achieving the necessary scale to make LLMs useful in medicine. As have been mentioned, access to large-scale, heterogeneous medical data sets where privacy and consent is assured is not a trivial matter. Once data is in place, there is also a considerable hardware cost associated with building such models—even though this is expected to decrease as technology improves¹³—and the carbon footprint to go along with it¹⁸. Recent research into smaller, more efficient models such as the work by Touvron et al. on LLaMA¹⁹ does however increase the likelihood of making LLMs more accessible for specific medical tasks.

4.4 HUMAN-AI COLLABORATION

While a lot of research on ML model performance in the clinical domain is of an adversarial nature, i.e., comparing human with AI performance²⁰, there is a growing realization that collaborative human-in-the-loop deployments may be a more realistic way of putting these models to practical use^{21,22}. As has been seen, it depends on the tested system whether AI/human cooperation is beneficial^{23,24} or not²⁵⁻²⁷.

It can be useful to draw parallels with the use of AI programming assistants such as GitHub Copilot, which are increasingly used by software developers as productivity tools. These assistants are perhaps the best current examples of LLM-based tools being used to support established process workflows. In a recent study by Perry et al., participants were told to solve a number of security related programming tasks with or without a code assistant²⁸. While the results suggested that developer productivity could increase, the authors also found that participants with access to the AI assistant tended to introduce more security vulnerabilities in their code than those without access. Moreover, the same participants were also more likely to believe that their code was secure. In effect, the AI introduced a false sense of security among less experienced users.

Another study by Vaithilingam et al. found that while developer productivity did not necessarily increase, the participants still preferred to use the AI assistant in their daily work because it saved them some effort in searching for information and getting started²⁹. However, difficulties with understanding the results produced by the assistant ended up significantly reducing their

effectiveness in solving tasks.

To make the most out of such assistants it may be necessary to understand more about how users interact with them, as suggested by Barke et al.³⁰ In their study they found that programmers interacted with the assistant in two modes: in *acceleration mode*, where the programmer knows what to do and wants it done faster, and in *exploration mode*, where the programmer is not sure how to proceed and needs help with exploring their options. A similar example can be found in a working paper by Dell'Acqua et al.³¹, which investigated the effect of AI support on consulting work. They found two distinctive patterns related to successful use of AI: while one set of consultants chose to delegate activities to either the AI or themselves, another decided to instead integrate their workflow completely with the AI and continuously interact with it.

It is likely that similar lessons around how e.g. the skill level, experience, preferences, and background information of the end user can influence both their interaction with the tool and how the generated answers are put to use will be another important factor for successful introduction of foundation—and other—models in healthcare decision-support.

4.5 THE ROAD AHEAD

The increasing capabilities of self-supervised foundation models has the potential to not only change how we approach clinical decision support systems but also healthcare information systems in general. Returning to the research questions, access to high-quality, unbiased data in a privacy- and consent-preserving manner remains a challenge, both in terms of providing sufficient volume for model training and finding ways of addressing bias and data quality issues (**RQ1**). On the other hand, the unsupervised nature of such models, combined with emerging techniques such as few-shot learning³², may reduce the need for data preparation and annotation. Moor et al.⁵ propose *generalist medical AI* (GMAI) as a possible paradigm for application of foundational models in medicine. If using a GMAI model trained on a diverse set of multimodal data as a backbone, adaptation to new tasks may take the form of model refinement through targeted prompting, reinforcement, and context adaptation.

This can have implications for the role of the healthcare professional and domain expert (**RQ2**), whose responsibility may shift from data annotation and knowledge modelling to that of prompting and reinforcement. Formal representations of medical knowledge are still useful; not only as input to foundation models but perhaps also for validation of their output. It remains to be seen how we constrain the unbridled creativity of large language models with our accumulated, evidence-based knowledge of what to do and what not

to do when administering healthcare.

Another open question is to what extent foundation models can make sense of patient history and trajectory data (**RQ3**). Their capability for effortless ingestion of multimodal input data (e.g. text, images, and laboratory results) opens up for deeper and more actionable representations of knowledge about the patient than before. Here too, there are concerns regarding how accurately and reliably such models can learn, recount, and reason over patient data. For example, we know that patient trajectory data will have gaps and omissions, but we may not want the neural network to fill them for us.

At the time of writing, it is still an open question as to how these new ways of querying and reasoning over data and knowledge will find their way into clinical practice. It may be a good opportunity to reconsider how we want to interact with such technologies, regardless of how and when eventual adoption will take place. The mixed results from existing implementations of CDSS, combined with our understanding of what it takes to make them work, indicate that improved algorithms and ML models alone may not be sufficient. As we learn more about the possibilities and limitations of the emerging generation of large language models, keeping an open mind about how they can enable new venues and modes of clinical decision support may be useful—as well as whether or not we have attempted to solve the correct problems with the most effective approaches at the right time. Toussaint³³ and Coiera³⁴ remind us that healthcare is a very communication-intensive practice. Perhaps exploring knowledge through natural language is a better way for healthcare professionals to interact with a CDSS. Chat may also serve as an interface that is more conducive to being used for exploratory, information-gathering tasks, obscuring the fact that the user is interacting with an advanced information system. Finally, it can also help improve the perceived agency of the end users, compared with e.g. the alert-based and form-filling approach of many current CDSSs.

The need to make informed, better decisions is becoming increasingly relevant in a healthcare domain with ever-expanding amounts of data and knowledge. New technologies have the potential to help us do so—but perhaps in different ways than we have assumed so far.

4.6 REFERENCES

- [1] Rishi Bommasani, *et al.* On the Opportunities and Risks of Foundation Models (2022). URL <https://arxiv.org/abs/2108.07258>. Cited on page/s 39, 40.
- [2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature* **521** (7553), 436–444 (May 2015). ISSN 1476-4687. doi: 10.1038/nature14539. URL <https://doi.org/10.1038/nature14539>. Cited on page/s 39.

- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems* volume 25. Curran Associates, Inc. (2012). URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. Cited on page/s 39.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17* pages 6000–6010 Red Hook, NY, USA (2017). Curran Associates Inc. ISBN 978-1-5108-6096-4. event-place: Long Beach, California, USA. Cited on page/s 39.
- [5] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature* **616** (7956), 259–265 (April 2023). ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4. URL <https://doi.org/10.1038/s41586-023-05881-4>. Cited on page/s 39, 42.
- [6] Tom B. Brown, *et al.* Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems NIPS'20* Vancouver, BC, Canada (2020). Curran Associates Inc. ISBN 978-1-71382-954-6. Cited on page/s 40.
- [7] Tiffany H. Kung, *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2** (2), e0000198 (February 2023). doi: 10.1371/journal.pdig.0000198. URL <https://doi.org/10.1371/journal.pdig.0000198>. Publisher: Public Library of Science. Cited on page/s 40.
- [8] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems (2023). URL <https://arxiv.org/abs/2303.13375>. Cited on page/s 40.
- [9] John W. Ayers, *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* **183** (6), 589–596 (June 2023). ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.1838. URL <https://doi.org/10.1001/jamainternmed.2023.1838>. Cited on page/s 40.
- [10] Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* **388** (13), 1233–1239 (March 2023). ISSN 0028-4793. doi: 10.1056/NEJMSr2214184. URL <https://doi.org/10.1056/NEJMSr2214184>. Publisher: Massachusetts Medical Society. Cited on page/s 40.
- [11] Karan Singhal, *et al.* Large Language Models Encode Clinical Knowledge (2022). URL <https://arxiv.org/abs/2212.13138>. Cited on page/s 40.
- [12] OpenAI. GPT-4 Technical Report. Technical report (2023). URL <https://arxiv.org/abs/2303.08774>. Cited on page/s 40.
- [13] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine* **29** (8), 1930–1940 (August 2023). ISSN 1546-170X. doi: 10.1038/s41591-023-02448-8. URL <https://doi.org/10.1038/s41591-023-02448-8>. Cited on page/s 40, 41.
- [14] Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K. Prasad, Adam Landman, Keith J Dreyer, and Marc D. Succi. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. *medRxiv* page 2023.02.21.23285886 (January 2023). doi: 10.1101/2023.02.21.23285886. URL <http://medrxiv.org/content/early/2023/02/26/2023.02.21.23285886.abstract>. Cited on page/s 40.
- [15] Anthony J. Nastasi, Katherine R. Courtright, Scott D. Halpern, and Gary E. Weissman. Does ChatGPT Provide Appropriate and Equitable Medical Advice?: A Vignette-Based, Clinical Evaluation Across Care Contexts. *medRxiv* page 2023.02.25.23286451 (January

- 2023). doi: 10.1101/2023.02.25.23286451. URL <http://medrxiv.org/content/early/2023/03/01/2023.02.25.23286451.abstract>. Cited on page/s 40.
- [16] Oded Nov, Nina Singh, and Devin M. Mann. Putting ChatGPT's Medical Advice to the (Turing) Test. *medRxiv* page 2023.01.23.23284735 (January 2023). doi: 10.1101/2023.01.23.23284735. URL <http://medrxiv.org/content/early/2023/01/24/2023.01.23.23284735.1.abstract>. Cited on page/s 40.
- [17] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* 6 (1), 135 (July 2023). ISSN 2398-6352. doi: 10.1038/s41746-023-00879-8. URL <https://doi.org/10.1038/s41746-023-00879-8>. Cited on page/s 41.
- [18] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink (2022). URL <https://arxiv.org/abs/2204.05149>. _eprint: 2204.05149. Cited on page/s 41.
- [19] Hugo Touvron, *et al.* LLaMA: Open and Efficient Foundation Language Models (2023). URL <https://arxiv.org/abs/2302.13971>. Cited on page/s 41.
- [20] Xiaoxuan Liu, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1 (6), e271–e297 (October 2019). ISSN 2589-7500. doi: 10.1016/S2589-7500(19)30123-2. URL [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2). Publisher: Elsevier. Cited on page/s 41.
- [21] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine* 28 (1), 31–38 (January 2022). ISSN 1546-170X. doi: 10.1038/s41591-021-01614-0. URL <https://doi.org/10.1038/s41591-021-01614-0>. Cited on page/s 41.
- [22] Bhavik N. Patel, *et al.* Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine* 2 (1), 111 (November 2019). ISSN 2398-6352. doi: 10.1038/s41746-019-0189-7. URL <https://doi.org/10.1038/s41746-019-0189-7>. Cited on page/s 41.
- [23] Ayush Jain, *et al.* Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Tele dermatology Practices. *JAMA Network Open* 4 (4), e217249–e217249 (April 2021). ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.7249. URL <https://doi.org/10.1001/jamanetworkopen.2021.7249>. Cited on page/s 41.
- [24] Ida Sim, Paul Gorman, Robert A. Greenes, R. Brian Haynes, Bonnie Kaplan, Harold Lehmann, and Paul C. Tang. Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *Journal of the American Medical Informatics Association* 8 (6), 527–534 (November 2001). ISSN 1067-5027. doi: 10.1136/jamia.2001.0080527. URL <https://doi.org/10.1136/jamia.2001.0080527>. Cited on page/s 41.
- [25] Pranav Rajpurkar, *et al.* CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine* 3 (1), 115 (September 2020). ISSN 2398-6352. doi: 10.1038/s41746-020-00322-2. URL <https://doi.org/10.1038/s41746-020-00322-2>. Cited on page/s 41.
- [26] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2 (3), e138–e148 (March 2020). ISSN 2589-7500. doi: 10.1016/S2589-7500(20)30003-0. URL [https://doi.org/10.1016/S2589-7500\(20\)30003-0](https://doi.org/10.1016/S2589-7500(20)30003-0). Publisher: Elsevier. Cited on page/s 41.

- [27] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Working Paper 31422 National Bureau of Economic Research (July 2023). URL <http://www.nber.org/papers/w31422>. Series: Working Paper Series. Cited on page/s 41.
- [28] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do Users Write More Insecure Code with AI Assistants? (2022). URL <https://arxiv.org/abs/2211.03622>. _eprint: 2211.03622. Cited on page/s 41.
- [29] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems CHI EA '22* New York, NY, USA (2022). Association for Computing Machinery. ISBN 978-1-4503-9156-6. doi: 10.1145/3491101.3519665. URL <https://doi.org/10.1145/3491101.3519665>. event-place: New Orleans, LA, USA. Cited on page/s 41.
- [30] Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proc. ACM Program. Lang.* 7 (OOPSLA1) (April 2023). doi: 10.1145/3586030. URL <https://doi.org/10.1145/3586030>. Place: New York, NY, USA Publisher: Association for Computing Machinery. Cited on page/s 42.
- [31] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (September 2023). URL <https://papers.ssrn.com/abstract=4573321>. Cited on page/s 42.
- [32] Archit Parnami and Minwoo Lee. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning (2022). URL <https://arxiv.org/abs/2203.04291>. Cited on page/s 42.
- [33] P.J. Toussaint and E. Coiera. Supporting communication in health care. *Supporting Communication in Health Care* 74 (10), 779–781 (October 2005). ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2005.04.007. URL <https://www.sciencedirect.com/science/article/pii/S1386505605000468>. Cited on page/s 43.
- [34] E. Coiera. When conversation is better than computation. *Journal of the American Medical Informatics Association : JAMIA* 7 (3), 277–286 (June 2000). ISSN 1067-5027 1527-974X. doi: 10.1136/jamia.2000.0070277. Place: England. Cited on page/s 43.

Part II

PUBLICATIONS

PAPER A

Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-Initiated Project

Development of a Medication Reconciliation Tool for Norwegian Primary Care EPR Systems: Experiences from a User-initiated Project

Thomas Brox Røst², Inger Dybdahl Sørby¹, and Gry Seland^{1,3}

¹ Vivit AS, Trondheim, Norway

² Atbrox AS, Trondheim, Norway

³ Gjøvik University College, Gjøvik, Norway

Abstract. Medication reconciliation is one of the most important priorities of national and international patient safety efforts, due to the numerous deaths and adverse drug reactions caused by inappropriate medication use. One of the main challenges of general practitioners (GPs) is to get an overview of changes in the patients' medications after transitions between healthcare institutions. This paper presents how Natural Language Processing of free text notes such as discharge summaries is used to automatically extract information about medications and how this can be compared to the patient's existing medication list in an electronic patient record (EPR) system. The functionality has been developed in a user initiated project, as a cooperation between four different vendors and with a strong involvement of the end users. The functionality is available for most Norwegian GPs and is seen as a very useful tool in the medication reconciliation process.

Keywords: Medication reconciliation, natural language processing, user centered development

1 Introduction

1.1 Medication Reconciliation

Medication reconciliation is the proposed formal, systematic strategy to overcome medication information communication challenges and reduce unintended medication discrepancies that occur at transitions in care [1]. When conducted as intended, medication reconciliation is a conscientious, patient-centred, inter-professional process that supports optimal medication management [2].

Copyright ©2014 by the paper's authors. Copying permitted for private and academic purposes.

In: E.A.A. Jaatun, E. Brooks, K. Berntsen, H. Gilstad, M. G. Jaatun (eds.): Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014), Trondheim, Norway, 19-MAY-2014, published at <http://ceur-ws.org>

54 T. B. Røst, I. D. Sørby and G. Seland

The lack of medication reconciliation is seen as a significant challenge to patient safety. Several studies have shown that different healthcare professionals, the patient and the relatives do not have the full overview of the patients' prescribed medications, particularly after transitions between different healthcare institutions. Not having the full overview of medication use is one of several causes of adverse drug reactions [3-5].

Several ongoing initiatives focus on processes where different health care providers such as physicians, nurses, and pharmacists cooperate with patients and their relatives to ensure accurate and consistent medication lists across transitions in care. In Norway, three of eleven focus areas in the Norwegian Patient Safety Programme: In Safe Hands¹ are related to medications: Medical reconciliation, drug review in nursing homes and drug review in home care services. Other programmes have been introduced in other countries. For example, the Institute for Safe Medication Practices in Canada² support medication reconciliation at a provincial, national and international level, and the Agency for Healthcare Research and Quality³ in the US has developed a toolkit for organizations to develop medication reconciliation based on knowledge of best practice.

However, the medication reconciliation process is tedious and time-consuming, and there has been a lack of electronic systems that facilitate the process of comparing and adjusting medication lists from different sources such as discharge summaries from hospitals or nursing homes and the "medications in use" list in the general practitioners (GPs) electronic patient record (EPR) system. In Norway, this has been done by the GPs, who had to print lists of medications on paper from their own EPR system as well as lists received in e.g. discharge letters, and then comparing each medication manually. The result then had to be entered into the EPR system. If the patient uses many medications the comparison process becomes complex and sometimes neglected.

This paper presents how natural language processing of free text notes such as discharge summaries has been used to automatically extract information about medications and how this can be compared to the patient's existing medication list in an EPR system. The functionality has been developed in a user initiated project, as a cooperation between four different vendors and with a strong involvement of representative end users.

1.2 Natural Language Processing

Several studies have shown how information technology and natural language processing (NLP) is used to facilitate the medication reconciliation process [6, 7]. Some of the main challenges are that the medication information is coming from multiple sources, using different controlled terminologies that has to be consolidated.

¹ Norwegian Patient Safety Programme, <http://www.pasientsikkerhetsprogrammet.no>

² Institute for Safe Medication Practices Canada, <http://www.ismp-canada.org/medrec/>

³ Agency for Healthcare Research and Quality, <http://www.ahrq.gov/qual/match/>

Basic lexical approaches, such as keyword matching, may sometimes be appropriate for detecting simple concepts from medical free-text [8, 9]. Problems such as understanding the medical context [10], recognizing negative terms and ending up with too many false positives [11] may, however, be common. To achieve higher accuracy, natural language processing techniques are often employed, at the cost of higher development complexity [12, 13]. The well-defined sub-language of the medical domain is often considered suitable for linguistic processing, given that the vocabulary is more restricted than in general language and sentences can be terse and to the point [14]. On the other hand, this creates its own set of problems when using e.g. parsers trained on typical corpora, such as newspaper texts, in particular related to ungrammatical language, spelling mistakes and non-standard abbreviations.

Traditional deep-linguistic grammars can be difficult to implement and are prone to producing too many and ambiguous results. A simpler approach is to use partial or shallow parsing [15]. With full parsing, the goal is to produce a complete parse tree of a sentence. A shallow parser will only concern itself with finding the parts of a sentence that are deemed relevant. While building a full parser for natural language will be a difficult task for even restricted medical domains, constructing a shallow parser is a far simpler option. Moreover, in many cases the problem is to identify the parts of a sentence that are of interest. With this in mind, shallow parsing can be a viable approach towards identifying medication administration events.

1.3 Project Overview

The project was initiated in 2011 by the Norwegian College of General Practice (Norsk forening for allmenmedisin - NFA) Reference Group for EPR. The members of the reference group are general practitioner specialists who have a special interest in ICT and EPR systems, and a particular interest in how to obtain improved systems that can function as useful tools for the GPs in their daily work with patients. Developing a tool for medication reconciliation was the top priority of the reference group, and a requirements specification for the solution was developed in cooperation with the Norwegian Centre for Informatics in Health and Social Care. The suggested solution was to develop a module for extracting and comparing medication information from different sources. During the period from September 2011 to September 2012, a project plan was developed and agreed between NFA, Vivit⁴, and the three major Norwegian general practice EPR system vendors. Vivit's role in the project was to develop functionality for recognizing and extracting medication information that could be integrated with

⁴ Vivit is a small Norwegian company with high competence and experience in the field of health informatics. Vivit focus on user-centered design and development, including empirical methods for user-centered requirements elicitation and analysis, usability testing and evaluation of clinical information systems, and methods for search, de-identification, and secondary use of electronic patient information. The company was founded in 2009 by former health informatics researchers from The Norwegian University of Science and Technology (NTNU).

56 T. B. Røst, I. D. Sørby and G. Seland

all the major Norwegian general practice EPR systems. The partners signed the project agreement in September 2012, and the first version of the system was implemented and tested by pilot users in February 2013. In May, 2013, version 1.0 was launched, and from January, 2014, the solution has been available for the majority of Norwegian GPs.

2 Objectives and Main Issues

The main objective of the project was to develop a method for extracting medication information from unstructured text.

Several issues had to be resolved to ensure a successful project outcome: The discharge notes that were to be matched against the medication information in the primary care EPR system were provided as free-text with no semantic or structural markup. Input from several different hospital and elderly care EPRs were to be expected, with no standardized ways of structuring the information. Accordingly, a method for extracting medication information from natural language had to be devised.

A prerequisite for the project was that the new medication reconciliation functionality had to work with three different EPR systems developed by separate, competitive, vendors. Each system had its own approach towards storing medication information, hence a joint interchange format had to be developed.

3 Methods

A set of 100 extracts from discharge notes were collected from various general practice patient records. The source material was discharge notes that were sent via electronic messaging from the hospital where the patient in question had undergone treatment. Since we were only interested in medication information, only the parts of the discharge note that contained such information were used. This also helped ensure that there was no identifying information in the extracted text. To evaluate the system, a gold standard was needed. An annotator was given the task of marking up all relevant medication information (medication name, dosage and frequency) in the 100 training notes. The annotation was done independently from the software development. The annotation was performed by a health informatics researcher. Based on the information available to us, we created a set of EBNF-like [16] grammars that represented various ways of describing medication information. The key part of the grammar were the terminals representing medication names. We made use of the FEST database, which is a national database containing information about all medications available on the Norwegian market. In addition to the medication names, FEST also includes associated ATC (Anatomical Therapeutic Chemical Classification System) codes, information about dosages, frequencies, and various ways of administering each drug. Much of this information could be imported directly as grammar terminals, thus simplifying the grammar building process. The grammars were compiled into a general text matching module, implemented in the

C# language. This module would take two inputs: The unstructured text (e.g. discharge notes) and a structured list of medications from the primary care patient record. The key steps of the processing pipeline were as follows: 1) The text would be split into sentences and tokenized; 2) text matching was applied, returning a list of extracted medications including their location, dosage and frequency; 3) the extracted medications were compared with the known medications, producing a list of matched extracted and known medications. The reconciliation between known and extracted medications was done by doing a combined semantic and syntactic comparison between the two medication sets. For instance, medications would be matched by both name and ATC code, meaning that two medications with different names but the same ATC code were eligible match candidates. Having found a set of possible match pairs, the additional information about dosage and frequency, including possible synonyms, would be included in the comparison. By calculating the Levenshtein string similarity metric [17] between normalized versions of all possible match candidates, the matches that resembled each other the most would be returned as match pairs. The remaining medications where no match was found would be returned as a single medication item with no corresponding match. Due to the use of string similarity measures, the match would be slightly fuzzy by nature, meaning that inexact matches were allowed. In practice, this turned out to be a minor issue, since the user interface could highlight differences between the matches. Also, the output from the extraction module was only intended as a decision support aid. Each suggested medication match would have to be approved manually with a conscious decision of whether or not the match was likely to be correct.

4 Results and Findings

Upon project delivery, the results were evaluated on a sample of 25 notes from the original 100 note test set. Table 1 summarizes the results. To understand the results, note that dosage and/or frequency will never be extracted if an associated medication is not found. If a medication has no dosage, it will still count as a true positive for MD and MDF if no dosage is extracted. Also note that no false positives were found for this evaluation. From user feedback we later learned of (and fixed) false positives, but in practice these are fairly rare.

Table 1. Evaluation Results

Match type	True positives	False negatives	Total	Precision	Recall
Medication (M)	201	10	211	100%	95,2%
Medication and dosage (MD)	175	36	211	100%	82,9%
Medication, dosage and frequency (MDF)	141	70	211	100%	66,8%

58 T. B. Røst, I. D. Sørby and G. Seland

4.1 Implementation in EPR Systems

The module has been implemented by the three EPR vendors that participated in the project. Figure 1 shows a part of a screenshot from the Infodoc Plenario EPR system ⁵.

Legemidler i bruk						Medikament fra tekst		
Vurderet	Kategori	Navn	Virkestoff	Sjykte	Dosering	Navn	Sjykte	Dosering
<input checked="" type="checkbox"/>	Fast	Albyl-E	Acetylsalisyls...	75mg	1 tbl dgl	Albyl-E	75mg	x 1
<input checked="" type="checkbox"/>	Fast	Orfiril	Valproinsyre	300mg	1 tbl 1 ggr dgl	Orfiril	300mg	x 1
<input checked="" type="checkbox"/>						Cipramil	10mg	10 mg
<input checked="" type="checkbox"/>						Zopiklon Mylan	5mg	
<input checked="" type="checkbox"/>	kur	Apocillin	Fenoksymetyl...	660mg	1 + 1 + 2	Apocillin	660mg	1 + 1 + 2
<input type="checkbox"/>						Lamotrigin	25mg	x 2

Inngående tekst til VIVIT fra Infodoc Plenario	Tekst etter tolkning av VIVIT
Cipramil Tab 10 mg 10mg Tablett - Blisterpakning Dssn 1 tbl dgl (Fast) Zopiklon Mylan 5mg Tablett - Blisterpakning Dssn 1 tbl kveld (Behov) Apocillin 660mg 1+1+2 (kur) Albyl-E 75mgx1 Orfiril tbl 300mgx1 (NY) Lamotrigin tbl 25mg x 2 (NY)	Cipramil Tab 10 mg 10 mg Tablett - Blisterpakning Dssn 1 tbl dgl (Fast) Zopiklon Mylan 5 mg Tablett - Blisterpakning Dssn 1 tbl kveld (Behov) Apocillin 660 mg 1 + 1 + 2 (kur) Albyl-E 75 mg x 1 Orfiril tbl 300 mg x 1 (NY) Lamotrigin tbl 25 mg x 2 (NY)

Fig. 1. Screenshot from the Infodoc EPR system

The lower left part of the figure shows the incoming text that has been pasted from a discharge summary. The lower right part of the figure shows the information that has been recognized by the Vivit module in bold text. The upper left part of the figure shows the initial medications in use list, and the upper right part of the figure shows the recognized medications from the discharge summary. The table is sorted in order to match similar medications on corresponding lines.

The physician has to assess every entry in the list in order to accept or reject the suggested changes of the list. New medications can easily be prescribed as the new values are automatically transferred to the prescription user interface.

4.2 The system in Use

All the vendors involved in the project have implemented the solution in their EPR systems, hence the solution is available for practically all Norwegian GPs. However, the vendors have different processes for making the new functionality known to their customers and not all GPs are aware of the functionality. In addition, the vendors have implemented the module in different ways, and the user interface solution may affect how the users perceive the usability of the medication reconciliation solution. A systematic evaluation of the solution is currently being carried out, but the results are not yet ready. However, the responses from many GPs are positive, although there is obviously room for improvement. Some GPs are enthusiastic and state that

⁵ <http://www.infodoc.no/>

The new tool for synchronizing medication lists is really simplifying the task of comparing and adjusting medication lists between our patient records and the hospitals. (Specialist General Practitioner, using the module with InfoDoc EPR system)

and

The new tool is very useful and makes my daily work easier! (GP, using the module with WinMed 3.0 EPR system)

Other users find the functionality useful, but miss more information about dosage and frequency, which is important when comparing medications. Further, the module is able to recognize and correct some spelling mistakes, but not all. If this could be improved the usability would have been better. At the moment, medication from unstructured text is better recognized than text in a semi-structured format, which is often used in the Care Sector in the Municipalities. Finally, some GPs would like the functionality to be more automatic, as it still may take considerable time to review the patients medications even with the module. After the different lists are compared, the physician must determine which medications the patient should use or not, and this has to be done for example by clicking on each individual medication.

5 Discussion

The new module does not make the medication reconciliation process automatic, but it offers a tool that can be seen as a decision support system that enables physicians to easily import medication information from various sources into the EPR system. The EPR systems were developed in the same programming language (C#), but they used different versions of C# and .NET. The developed functionality had to work with all systems, so a lowest common denominator approach had to be used when writing the software. This also put some constraints on the use of third-part libraries.

There is no standardised way of denoting medication information. As an example, when denoting medication frequency, the terms x 1 and [1+0+0+0] both mean the same thing (once a day). To some extent the project had to cope with such syntactic differences.

As with almost all clinical text, spelling errors are common in discharge notes. It was a strong requirement that minor spelling mistakes should be handled. We had access to 100 discharge notes as training/example data. This is a fairly small amount of data, which made the use of machine learning methods difficult. Moreover, no gold standard was available, meaning that this had to be developed as part of the project. The annotation was performed by a health informatics researcher with background in computer science. Using more annotators with different backgrounds (e.g. pharmacists, healthcare professionals) could probably have increased the validity of the annotation.

The recall rate for medications is mostly explained by the use of medication names that were not found in the FEST database (typically colloquial terms)

60 T. B. Røst, I. D. Sørby and G. Seland

and major spelling errors that the module was not able to correct. The lower recall for frequency descriptions can be ascribed to a larger variety of expressions when describing frequencies than what is the case for dosages. The evaluation of the module shows that most medications are recognized, while the recall for dosage and frequency is lower. However, as it is easy to change dosage and frequency values, the users find the functionality very helpful as long as most of the medications are recognized. As more end-users start using the new functionality, feedback and error reports are basis for continuously improvement of the grammar. With partial parsing, the grammar strictly defines the elements that we are able to extract. This means that every false positive requires adding additional rules to the grammar. To make this work, the parser developer must take care so that grammar additions do not break previous functionality. In our experience this calls for a structured, iterative approach to grammar development. Having a full set of unit test cases makes grammar development a lot easier and safer. A positive side-effect is that with this approach the precision is usually very high, at the expense of lower recall. Another problem with shallow parsing approach is that it is not ideal for extracting complex narrative. For simple medication, dosage and frequency extraction we have seen that this is not a big problem due to the usually structured notation. However, building a grammar that e.g. extracts the reasoning behind a prescribed medication will be a lot more difficult, this because reasons usually are given in natural language. A shortcoming with the evaluation was that it was performed on the same material as was used for building the grammars, this due to the relatively small amount of data available to us.

5.1 User Involvement

Experiences from several ICT projects in the health care sector show that a lot of projects do not involve end users in the development. The results are often systems that can be found annoying and time-consuming, failing to meet the needs of the end users. The project presented in this paper was initiated by highly active and engaged users with a real need for improved functionality of their EPR systems. The involvement of the users and their participation in the pilot testing and the approval of the solution was a clear advantage in order to ensure the success of the project.

6 Conclusions and Further Work

The project presented in this paper has shown how joint efforts and cooperation of end users and vendors have led to the development of new, common functionality that supports the medication reconciliation process.

6.1 Further Work

As part of the project follow-up, the medication reconciliation module will be continuously improved and updated bi-monthly with the latest version of FEST as well as general quality improvement. This will ensure that the module stays up to date.

Repeating the evaluation on a new set of test data will give a clearer view of real-world precision and recall. Another important aspect is to measure end-user satisfaction with the medication reconciliation tool and how this affects their daily work. For this purpose, a survey is being performed on end users and results are likely to be ready later this year.

To improve recall, making use of machine learning technologies is a viable option. This will, however, require more test data. A benefit of having a hand-crafted extraction module with high precision is that it can probably be used for automated annotation (i.e. applying the module to new test data) so as to make the gold standard creation job much easier.

List of Abbreviations Used in the Paper

ATC	Anatomical Therapeutic Chemical Classification System
EBNF	Extended Backus-Naur Form
EPR/EHR	Electronic Patient Record/Electronic Health Record
FEST	Prescription and expedition support (Norwegian: Forskrivnings- og ekspedisjonsstøtte)
GP	General Practitioner
ICT	Information and Communication Technology
NLP	Natural Language Processing

62 T. B. Røst, I. D. Sørby and G. Seland

References

1. Lo, L., Kwan, J., Fernandes, O.A., Shojania, K.G.: Medication reconciliation supported by clinical pharmacists. In: Making Health Care Safer II: An Updated Critical Analysis of the Evidence for Patient Safety Practices. Evidence Reports/Technology Assessments, No. 211. Rockville (MD): Agency for Healthcare Research and Quality (US) (2013)
2. Greenwald, J., Halasyamani, L., Greene, J., LaCivita, C., Stucky, E., Benjamin, B., et al.: Making inpatient medication reconciliation patient centered, clinically relevant and implementable: A consensus statement on key principles and necessary first steps. *Journal of Hospital Medicine* **5**(8) (2010) 477–85
3. Mjörndal, T., Boman, M.D., Hägg, S., et al.: Adverse drug reactions as a cause for admissions to a department of internal medicine. *Pharmacoepidemiol Drug Saf.* **11**(1) (Jan-Feb 2002) 65–72
4. J. E., I. B., J. E., et al.: Drug-related deaths in a department of internal medicine. *Archives of Internal Medicine* **161**(19) (2001) 2317–2323
5. DT, L., JD, K., DB, P., A, L., RG, B.: Hospitalization and death associated with potentially inappropriate medication prescriptions among elderly nursing home residents. *Archives of Internal Medicine* **165**(1) (2005) 68–74
6. Bassi, J., Lau, F., Bardal, S.: Use of information technology in medication reconciliation: a scoping review. *Annals of Pharmacotherapy* **44**(5) (2010) 885–897
7. Cimino, J.J., Bright, T.J., Li, J.: Medication reconciliation using natural language processing and controlled terminologies. In: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, IOS Press (2007) 679–683
8. Giuse, D.A., Mickish, A.: Increasing the availability of the computerized patient record. *Proc AMIA Annu Fall Symp* (1996) 633–7
9. Goldman, J.A., et al.: Term domain distribution analysis: a data mining tool for text databases. *Methods Inf Med* **38**(2) (1996) 96–101
10. Honigman, B., et al.: A computerized method for identifying incidents associated with adverse drug events in outpatients. *Int J Med Inform* **61**(1) (2001) 21–32
11. Murff, H.J., et al.: Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc* **10**(4) (2003) 339–50
12. Hripcsak, G., et al.: Unlocking clinical data from narrative reports: A study of natural language processing. *Ann Intern Med* **122**(9) (1995) 681–688
13. Hripcsak, G., Kuperman, G., Friedman, C.: Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of Information in Medicine* **37**(1) (1998) 1–7
14. Spyns, P.: Natural language processing in medicine: an overview. *Methods Inf Med* **35**(4-5) (1996) 285–301
15. Li, X., Roth., D.: Exploring evidence for shallow parsing. *Proc of Annual Conference on Computational Natural Language Learning* (2001)
16. Scowen, R.S.: Extended bnf – a generic base standard. *Software Engineering Standards Symposium* (1993)
17. Navarro, G.: A guided tour to approximated string matching. *ACM Computing Surveys* **33**(1) (2001) 31–88

PAPER B

Identifying Catheter-Related Events Through Sentence Classification

Identifying Catheter-Related Events Through Sentence Classification

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby,
Ingrid Andås Berg, Øystein Nytrø

June 22, 2020

Contents

1	Abstract	2
2	Introduction	2
3	Objectives	3
4	Related work	4
5	Methods	5
5.1	Data	5
5.2	Annotation	8
5.3	Data Analysis	10
6	Experiments	14
7	Conclusion	19
8	Acknowledgements	20
9	Figures	21
10	Bibliography	24

1 Abstract

Infections caused by central venous catheter (CVC) use is a serious and under-reported problem in healthcare. The CVC is almost ubiquitous in critical care because it enables fast circulatory monitoring and central administration of medication and nutrition. However, the CVC exposes the patient to a risk of blood-stream infections (BSI). Explicit documentation of normal CVC usage and exposure is sparse and indirect in the health record. For a clinician, CVC presence is simple to infer from record statements about procedures, plans and results related to CVC. In order to capture evidence about CVC-related risk of infections and complications, it is important to develop computerized tools that can estimate individual patient days of CVC exposure retrospectively for large cohorts of patients. Towards that objective, we have developed methods for learning classifiers for statements about CVC-related events occurring in the textual health record. This includes developing and testing an annotation ontology of events and indicators, annotation guidelines, a gold standard of annotated clinical records selected from a corpus of complete health records for more 800 episodes of care and collecting alternate health register evidence for validation purposes. This paper describes the available data and gold standard, feature selection approaches and our experiments with different classification algorithms. We find that even with limited data it is possible to build reasonably accurate sentence classifiers for the most important events. We also find that making use of document meta information helps improve classification quality by providing additional context to a sentence. Finally, we outline some strategies on using our results for future analysis and reasoning about CVC usage intervals and CVC exposure over individual patient trajectories.

2 Introduction

The use of intravenous cannulation is a very common procedure when a patient is treated in a hospital setting. The most common type of cannulation is performed with peripheral venous catheters (PVCs), since peripheral veins are readily accessible for insertion (Mermel, 2017). Central venous catheters (CVC) are primarily used to administer medications and fluids and to measure central venous pressure (Taylor and Palagiri, 2007). They typically consist of a tube that is inserted into one of the central veins of a patient. How long a patient is in need of a CVC varies from a couple of days to several months. The use of CVC in medical treatment is indispensable and life-saving for many patients but also exposes them for risk of infec-

tion and consequently increased morbidity and mortality (McKibben et al., 2005). Bacteria that are colonised on the catheter may cause a catheter-related bloodstream infection (CRBSI). For the first 3-4 days of CVC usage the risk is low (Fletcher, 2005). As the number of CVC usage days increases, so does the risk of CRBSI. This is a severe complication of CVC usage and may lead to hospital-acquired sepsis and in worst case death. More than 15 % of patients experience one or more complications during CVC insertion or maintenance (Taylor and Palagiri, 2007). Common complications in addition to catheter-related infections include arterial puncture, hematoma, pneumothorax, and venous thrombosis. Of these, catheter-related infections and venous thrombosis are often deadly. In some cases the mortality rate may be as high as 25 % (Brun-Buisson, 2001). Even though CVC usage is common we do not know enough about the prevalence and duration of CVC use, CVC-related infections and the associated patient injuries (Wong et al., 2018).

CVC-related infections are risky for the affected patient and costly to treat, often leading to prolonged hospitalization. A 2008 study of CRBSI in an intensive care unit found that each CRBSI event added approximately USD 82,000 in extra costs and 14 additional hospital days (Cohen et al., 2010). In a 2002 study of healthcare-associated infections in U.S. hospitals, the highest death rates were associated with bloodstream infections in intensive care units. Of a total of 81,942 infections, 25 % of these had death as an outcome.

Surveillance regimes and adverse event detection are the preferred approaches to increased quality of care and is mainly performed in intensive care units. These regimes require considerable manual labor, do not give much clinical effect, and may not be applicable in all hospital wards. In Norway, quarterly prevalence surveys are used to describe the current state of all hospitalized patients, but are not sufficient for estimating risk related to days of CVC usage. Ideally, we would like to use retrospective patient data to derive a precise risk ratio of CRBSI per CVC-day, and thus gain more detailed knowledge about an important patient safety indicator. In turn, this can guide better practice related to central-line catheter usage.

3 Objectives

In this paper we describe our research on automated retrospective capture of CVC-related events from a data set of annotated clinical notes. The project was performed in collaboration with researchers at Akershus University Hospital (Ahus). From their experience, there is insufficient knowledge about

prevalence and duration of CVC use for patients in Norwegian hospitals. The duration of CVC use (number of CVC days) is an important prerequisite for estimating the risk of CRBSI, and a first step towards targeted quality improvement work. It is also desirable to have better data on CVC insertion and removal events, without relying on explicit coding.

Our approach was to manually annotate the content of clinical notes with CVC-related events and states and then train machine learning classifiers on the annotated data set. Identifying events such as CVC insertion, care and removal can contribute to a faster and more accessible overview of the occurrence and duration of CVC usage. It can also provide improved monitoring of CVC-related bloodstream infections, thus contributing to patient safety. Moreover, detecting CVC placement can also be of use when performing risk evaluations.

To our knowledge, using machine learning and natural language processing for detecting CVC-related events has not been done previously on clinical notes in the Norwegian language. The work of Penz et al. (2007) on English-language clinical notes is similar but relies on a semi-automated approach and was targeted towards adverse events. Our focus is on CVC exposure time in general, and more specifically individualized risk assessment. This CVC-specific work is part of more general research on capturing episodes and exposure in health records.

4 Related work

In a systematic review of 200 studies related to bloodstream infections and intravascular devices, Maki et al. (2006) found that CVCs have far higher incidence rates than peripheral intravenous catheters and midline catheters. A study by Hojsak et al. (2012) investigated the rate of CVC-related sepsis for patients on parenteral nutrition. They found that CVC was used on average 243.9 days per patient. Because of septic episodes 12.8 % of the used catheters were removed. The importance of intervention and monitoring of catheter use was demonstrated by Pronovost et al. (2010). For a total of 300,310 catheter days their Keystone ICU quality improvement project saw a mean and median reduction of CRBSI from 7.7 and 2.7 to 1.3 and 0 over a 16-18 month period. Bruin et al. (2012) applied a fuzzy logic-based system to generate rules for early detection of CVC-related infections. Trick et al. (2003) evaluated the ability of the SymText natural language processing (NLP) system to find mentions of CVC in chest radiograph reports. SymText yielded a sensitivity of 95.8 % and a specificity of 98.7 % when compared with human interpretation. Penz et al. (2007) compared the per-

formance of an NLP program (MedLEE) and a phrase-matching algorithm in detecting CVC-related adverse events from clinical records. They found that phrase matching was a sensitive but non-specific method while the NLP program was less sensitive but significantly more specific. Combining the methods gave an acceptable sensitivity (72.0 %) and specificity (80.1 %). Another interesting finding was that incomplete or inaccurate clinical notes hampered all methods, including manual chart review. In a 2014 study by Michelson et al. (2014), text mining methods were found to be very effective in detecting different types of surgical site infections (SSIs). They did not consider CRBSIs specifically but their system was able to identify 100 % of infection cases detected by regular surveillance as well as 37 cases not previously identified. Bates et al. (2003) and Govindan et al. (2010) both give comprehensive overviews of various adverse event detection approaches. In general, there is pervasive research on retrospective NLP analysis of health data for many other purposes, though this falls outside of the scope of this paper.

5 Methods

5.1 Data

To perform our experiments we needed to build a dataset from a cohort of patients with CVC exposure. Our study design specified that we needed the patient records for both patients with CVC use as well as patients with both CVC and BSI. Following that we would acquire a larger set of reasonably similar patient records where CVC may or may not have been used. Rather than extracting the full patient record for a given patient we narrowed the data requirements down to a single episode of care. Our definition of an episode of care corresponds to the one used in the Norwegian specialist healthcare patient register NPR ¹. For an episode of care, we acquired the continuous set of clinical notes from the hospital. The patient may have had other contacts with healthcare providers, documented in separate records, but we did not collect those records from the same period. All episodes were to be selected from the DIPS EHR database of Akershus University Hospital (Ahus). Ahus is a tertiary-level university hospital that often receives and transfers patients to other hospitals. By only selecting complete episodes, i.e.

¹An episode of care is a period where the patient receives care and treatment from one institution for one health problem. An episode may be an outpatient visit, a day visit or a hospitalization, potentially with interspersed leaves. An episode designates activity, not only treatment.

including a concrete initial admission and final discharge, we largely avoided truncated episodes.

We decided to only extract the text in the clinical notes themselves and not any accompanying structured data. One reason was that we did not consider any of the available structured information directly useful for our purposes. There are specific NCSP (NOMESKO Classification of Surgical Procedures) surgical codes, such as PYGC00 ("Insertion of central venous catheter") and related codes, that we could use both for corpus selection and as a classification feature. This code could have been recorded in the structured part of the EHR and then possibly reported to national registries. However, when searching for this and related codes in the EHR, the number of results returned was much too low to be realistic. This was not unexpected: Part of the rationale for this study was that structured reporting of CVC use was lacking. Generally, ubiquitous procedures are not counted or documented separately if they are obvious or implicit in more comprehensive procedures. Moreover, reimbursement calculation (DRG coding) for intensive care patients tries to model severity and complexity, and CVC usage is not a distinguishing feature. In addition, our view of the patient was somewhat limited. Many patients would be transferred to or from the hospital which meant that the relevant surgical coding may not have been visible in the records available to us. The lack of CVC-related structured coding is also known from other research. In a paper on CVC adverse event detection, Penz et al. (2007) found that the unstructured text was the best source for finding patients with CVC.

Given that the coding could not aid corpus selection, we decided to make use of prevalence surveys instead. In Norway, all hospitals must perform two annual surveys on infections and the use of antibiotics. In such a survey the CVC state, in addition to several other parameters such as known infections, is recorded manually at a given date and time. Fortunately, this is done four times a year at Ahus and we thus decided to base our corpus selection on patients present on one of the four survey dates (Løwer et al., 2013) at the hospital. We needed a corpus containing a sufficiently large number of patients with CVC and decided on the following selection criteria:

1. For six quarterly prevalence survey days, all health record notes for the ongoing episode, for all patients registered as having CVC on the prevalence survey day were extracted. The identity of episodes or patients, or actual survey findings, were unknown to researchers and not represented explicitly in the record.
2. For a seventh prevalence survey day, complete episode health record notes for all inpatients in the most relevant departments were included.

This was to give us a representative set of similar patients, not necessarily having CVC at the prevalence survey date. Still, these patients could be expected to have many similar traits and findings and be subject to (peripheral and urinary) catheters.

We required the episode length to be at least four days. The rationale for this lower bound on episode length was to increase the total volume of the corpus. Some episodes spanned more than one prevalence surveys, but duplicates were removed in the final corpus. We could not identify if unique patients gave rise to more than one distinct episode but this was irrelevant for our study. Following this approach we ended up with a corpus which is summarized in Table 1.

Table 1: Corpus overview: Episodes and notes

Survey	Episodes	Notes	Notes	
			² Inspected	Annotated
1	44	2708	2708	377
2	28	2883	2883	432
3	14	1369	1369	165
4	23	1595	1595	190
5	57	2808	2804	341
6	22	2147	2147	289
7	631	³ 32104	⁴ 8668	⁵ 951
Totals	⁶ 819	45614	22174	2745

As mentioned we would extract all clinical notes available to us for each selected episode of care, including nursing notes, surgical notes, physician notes, laboratory examinations, and more. The average number of clinical notes for each patient was high enough to ensure that at good selection of notes both with and without CVC use were included. Considerable effort was needed both to retrieve, organize and clean the data, as described by Husby (2014) and Berg (2014). Each note in our corpus was represented as a plain-text file. Since all notes in the EHR were originally in RTF format they needed to be converted to a plain-text format without losing any formatting that was relevant to the interpretation of the note. The EHR vendor did not provide built-in conversion to text so a custom solution had to be built.

²Read, but not annotated

³All manually searched for content potentially relevant for CVC

⁴Positive search results, manually inspected

⁵True positives

⁶Some episodes are counted more than once, because they last longer than 3 months

The initial corpus would contain personally identifying information (PII) about both patients, staff, family, and other related people. This meant that the research would sort under the Norwegian Medical and Health Research Act (hfo, 2008), which stipulates that the Norwegian Regional Committees of Medical Research Ethics (REK) had to be involved. The research plan and objectives, including descriptions on how PII would be handled, was submitted to the committee, which then evaluated the research ethics of the project and finally approved our application. The application stated that only named researchers in the EVICARE project who had signed non-disclosure agreements would have data access. The data would be stored on an offline restricted local network where all access would be logged with timestamps and the identification of the accessing researcher. The physical server was only accessible to system administrators.

5.2 Annotation

To identify the clinical state documented by the clinical notes in our corpus, we defined a set of CVC-related annotation labels (Table 2). This was done as a collaboration between the authors and a domain expert in natural language processing. The classes of patient states labelled were intended to form a generalization hierarchy, e.g. "CVC" being a more general type of CVC-state than "Hickmann". When applied to the text, the annotator would label one or more words that would (roughly) act as a confirming proof of a certain state, situation or event. In practice, this meant that an annotation could span everything from a single word to a complete sentence.

The classes were intended to form an ontology about events, states, devices, conditions and symptoms. However, sparsity of events and non-documented care for CVC skewed our results. Furthermore, it was a continuing challenge to separate clinically implicit patient state from textually explicit record statements when assigning labels. I.e., what a trained clinician would be able to infer about patient reality and what could be read in the text documents. For the purpose of identifying CVC-state, we had to re-interpret the labels, and this is further discussed in section 5.3.

All the notes were translated to plain text, retaining sections, section headings and cleaning punctuation and sentence-dividers; see section 6 for details. Each note was saved into a single file. Each file was named with a unique serial number, patient ID, episode of care ID, the note type, and a timestamp showing when the original clinical note was written. No other correction or parsing was applied, so the individual note would have the appearance of a well formatted clinical note. The Brat rapid annotation tool (Stenetorp et al., 2012) was set up with the designed annotation ontology,

accessing one file at a time. Some test annotations were done during ontology design, but this was discarded once the annotation guideline was agreed upon and considered stable. The annotator, which is also one of the authors of this paper, was a nurse with special competence in infection control. For each processed clinical note file a corresponding annotation file was created. For each annotation, this file had a line with a local identifier, the annotation label, the start and stop character for the annotated text (referring to the original note), and the text fragment from the original note that was annotated. The following example shows what a single annotation could look like:

T1 RemCVC 241 277 CVC removed and tip sent for culture

If no annotations were made, an empty annotation file was still created; this would tell us that the file had been reviewed by the annotator but was without any annotated findings. The annotation files were named identically as the corresponding note, but given a different suffix. After the annotation process was completed we had a total of 22,174 notes. All the notes from survey days 1 to 6 were annotated. Only a quarter of the notes included after the day 7 prevalence survey was annotated. This fraction was determined by time and resources available after annotating all the notes for the episodes included because of the other 6 survey days (Table 1).

Table 2: Annotations

Annotation	Description
Carecvc	Care, observation or assessment of CVC.
PlanCarecvc	Care of CVC has not been performed, but has been booked or planned.
PlanInscvc	Admission of CVC not performed, but planned, desired or ordered for the future.
Inscvc	CVC has been inserted in the period covered by this note.
Remcvc	CVC has been removed in the period covered by this note.
PlanRemvcvc	Removal of CVC has been planned or ordered for the future.

Continued on next page

Continued from previous page

Annotation	Description
Symptom	Statements indicating that there may be a blood system infection (BSI).
Sepsis	Sentence containing the word "sepsis" or mention of similar conditions.
CVC, Hickmann, VAP, other	Labels for more or less specific type of CVC.
JugularVein, SubclavianVein, Femoralis	Labels for site of CVC.
Possiblecvc	Sentences where CVC is discussed without implication that CVC is present.

5.3 Data Analysis

Some of the note types, in particular the nursing notes, had a distinct format. This reflected the document editing interface in the EHR system, which came with predefined templates to structure the documentation process. The nursing notes often used a template with 12 different headings. Example headings are "Communication/Senses", "Breathing/Circulation", "Pain/Sleep/Rest/Well-being", and "Skin/Tissue/Wounds". Most nursing notes would follow this template, but typically only a subset of the sections would be used. For the most frequently occurring note types where such a structure existed we built regular expression-based parsers to extract the contextual information along with the text. The assumption was that knowledge about the context of a piece of text could potentially be used as a feature to enhance its interpretation. We also knew from Husby (2014) that approximately 10 % of the nursing notes would contain CVC-related information under the "Skin/Tissue/Wounds" heading, thus making the section information a potentially valuable feature. For this project we chose not to apply any deeper linguistic analysis, such as e.g. part-of-speech analysis. We had previous experience that clinical language was often terse and grammatically incomplete. Furthermore, we did not have access to comprehensive vocabularies of clinical terms for entity recognition.

Once the parsers had been tested and refined sufficiently there were still 65 out of the original 45,614 notes that would not pass, usually because the structure had for some reason been mangled. Given the total volume of clin-

ical notes we decided it was safe to discard these. We also chose to discard 1,892 notes that we thought were not relevant because of their note type. Examples of these were letters to the patient or to other healthcare institutions. Following this we were left with 43,657 notes. The final reduction of the corpus was to remove duplicated notes. From our initial data analysis we observed that several notes were exact duplicates where only the timestamp differed. Discussions with technical staff revealed that this was an artifact of how the EHR worked: Whenever a clinical note was reopened a new note would be generated, even if no changes were made. This could happen when a nurse opens a document for editing, but only read it. A similar situation of semi-duplication occurs when a document is edited incrementally. This creates a new note only slightly different from the previous one. We observed some cases where this happened but did not do any analysis of how prevalent this was; this could be relevant for future work. After removing duplicates, the final corpus size was 42,806 clinical notes. Another corpus reduction task we considered was to discard notes with infrequently occurring note types. Ultimately we decided against this as it would potentially affect the episode of care length.

Once we had extracted the text along with associated meta information we grouped all the note data according to the episode of care. The final processed corpus contained 778 episodes of care with 122 different types of clinical notes. Table 3 shows the most frequently occurring note types. The nursing notes were by far the most common note types. This made sense given that nurses are working three shifts and have a need to communicate throughout the day for continuity of care. For 50 of the 122 note types there were less than 5 note examples, making this a fairly long-tailed distribution. While e.g. somatic nursing notes are subdivided into "care", "plan" and "evaluation", the table aggregates this type for compactness. However, we treated these different nursing notes as separate in the analysis.

As shown in Table 1, approximately 50 % of the notes were inspected, of which 2,745 received annotations. The 10 most annotated note types are shown in Table 4. The rightmost column shows the number of notes where actual CVC annotations were made, i.e. not just empty annotation files. Of the remaining notes, 4,056 were read and 564 had annotations.

Table 5 shows how the annotation classes are distributed over the annotated notes. The most common class is CVC care (including observation and assessment), which makes sense given that this is an action likely to be performed during a nurse visit. Note that the number of CVC insertion and removal annotations differ. This can be explained by the CVC already being present when the patient arrives at the hospital or not being removed before leaving or being transferred. It may also be the case that documentation

Table 3: Note types, translated

Note Type	Count
Somatic nurse note (care, plan, evaluation)	28265
Somatic physician note	6641
Intensive nurse note (care, plan, evaluation)	1830
Somatic physician discharge summary	727
Somatic nurse ward admission note	596
Somatic medical admission note	574
Somatic nurse ward transfer note	426
Somatic nurse reception note	415
Somatic nurse summary	305
Somatic physician discharge note	183

Table 4: Annotated note types, translated

Annotated Note Type	Total	Annotated
Somatic nurse note (care, plan, evaluation)	2942	380
Somatic physician note	660	105
Intensive nurse note (care, plan, evaluation)	137	16
Somatic nurse ward transfer note	51	2
Somatic nurse ward admission note	18	4
Somatic physician discharge summary	17	8
Somatic medical admission note	16	4
(Somatic, physician) Transfer note	16	3
Palliative note	16	0
Somatic nurse ward admission note	14	5

is missing or incomplete, although this is less likely given the seriousness of the procedure. Another possibility is that the CVC spans more than one episode of care. A further complication is that more than one CVC may be present—we found cases of up to three CVCs being present—and inserted at different times, but removed together.

Table 5: Annotation count

Annotation	Description
Carecvc	349
Symptom	123
PlanInscvc	82
Inscvc	63
PlanCarecvc	54
Remcvc	50
CVC	37
Possiblecvc	35
Sepsis	32
PlanRemvcvc	22
JugularVein	19
Hickman	13
SubclavianVein	6

In Figure 1 we see the distribution of the number of notes in each episode of care. The mean number is 55 while the median is 34. The longest episode of care in terms of the number of notes had 643 notes. The similar statistic for episode of care duration in number of hospitalization days is shown in Figure 2. Here the mean was 29 days and the median 13. This reflects all the episodes which involve patients with CVC.

After inspecting some of the longest episodes it turned out that there were mostly sound medical reasons behind the long hospitalizations. In many ways this was expected, given that CVC use is often associated with serious medical conditions. There were, however, exceptions. In the episode with the longest duration, which lasted 361 days, it turned out that the actual admission period was approximately a fortnight. Almost a year after the discharge a single clinical note was tacked onto the episode, containing a standardized report to the national cancer registry. These deviations were also likely to occur for other episodes, so some care was needed if the admission period was to be used as e.g. a feature. We decided not to consider this as a problem since the episode length was not used in our experiments.

Table 4 shows that most of the annotation events are very sparse. This

was a challenge, given that sparse classes is a common problem in machine learning. To alleviate this we decided to make use of the intended generalization hierarchy of annotation event classes. In terms of semantics, classes such as CVC care and use are fundamentally quite similar, meaning it is probably safe to group them together into a common class. Besides, for our research purposes it was not necessary to exactly predict the given annotation labels: Our interest was in the CVC usage prevalence and duration, which means that the main goal was to detect the transitions between having and not having CVC. Accordingly, we decided to create four aggregate classes from the initial fifteen: *Plan* (PlanInscvc), *Ins* (Inscvc), *Use* (Carecvc, PlanCarecvc, CVC, PossibleCVC, PlanRemcvc, JugularVein, Hickman, SubclavianVein) and *Rem* (Remcvc). Note that planning removal of a CVC implies that the CVC is present. The reasoning behind these classes were that they should be sufficient to support our future attempts to infer periods of continuous CVC use. Note that the Sepsis and Symptom classes were discarded for now, even though they represent a substantial number of annotations. This was done because the Sepsis and Symptom labels were often used in situations that were unrelated to actual CVC use and could as such be a source of confusion to the classifiers. Table 6 shows the final distribution of our new aggregate classes. There is still some imbalance although to a smaller extent than before. As expected, the majority of samples are in the *Use* class.

Table 6: Aggregate annotation count

Note type	Count
Plan	82
Ins	63
Use	535
Rem	50

Finally, Figure 3 shows the aggregate annotation class frequency relative to the most common note types and Figure 4 shows the same information relative to the different sections in the somatic nursing notes (excluding sections without annotations). The numbers in parentheses show the total number of observations. For some documents and sections some of the sparser aggregate classes occur with a relatively high frequency.

6 Experiments

In order to find evidence of CVC use we decided to build text classifiers that would, given clinical notes as input, make predictions as to whether or not

one of our previously mentioned aggregate annotation classes should apply. In practice, the output classes would then be no CVC use (*None*), CVC planning (*Plan*), CVC insertion (*Ins*), CVC use (*Use*), and CVC removal (*Rem*). This would give us a foundation for later prediction of CVC usage intervals. Rather than classifying the whole note, we instead opted for classifying sentences given that the annotations were granular enough to attribute them to a particular sentence. This would also make it easier to use section information as an additional feature.

Our tool of choice for cleaning up the section notes and converting them into sentences was the Python NLTK Natural Language Toolkit (Loper and Bird, 2002). To perform sentence splitting with sufficient quality we used the NLTK Punkt Sentence Tokenizer. This tokenizer could be trained with our clinical notes as input data to perform unsupervised sentence boundary detection (Kiss and Strunk, 2006). We found that it was easily confused by abbreviations and spelling errors, both of which are common in clinical notes. To alleviate this we had to manually add said errors and abbreviations to the tokenizer, thus gradually improving its quality. After several iterations of manual review and corrections we found that the tokenizer yielded sufficient although not perfect quality on our source material. The fact that the source language was Norwegian did not pose any problems, so no translation or other modifications was necessary for the sentence splitting to work as intended.

As mentioned, the nursing notes largely followed templates with fixed section headers, author roles, hospital department and other information. We were particularly interested in the section information but also the other available information. After extracting the sentences from each note the resulting information was placed into a JSON data structure where each sentence was associated with relevant meta information, including the section header. If no section header information was available, as was the case for notes other than nursing notes, the sentences were given a "general" section header label.

To train our sentence classifiers we chose to use the Python scikit-learn library (Pedregosa et al., 2012). This is a well-established and efficient machine learning and data analysis toolkit which provided the functionality we required for this experiment. Data pre-processing yielded a total of 344,563 sentences. From these we selected 34,810 sentences that had been through the annotation process, out of which 640 had actual annotations. From a machine learning point of view this can be considered a fairly small data set, so we decided on using 4-fold cross validation rather than the more common 10-fold approach. Given the highly imbalanced data set (most sentences belonged to the *None* class) we considered whether or not stratification would make sense. Experiments both with and without fold stratification indicated

that stratified folds slightly alleviated the class imbalance problem and provided overall better classification performance. Accordingly, we settled on stratified folds. Our task was a multiclass classification problem and we decided to take the one-versus-all approach in this experiment.

Using the scikit `TfidfVectorizer` we gave each sentence in the training data set a tf-idf representation, using sublinear tf scaling and a `max_df` parameter setting of 0.5. This choice would remove frequently occurring words and was an alternative to techniques such as removing stop words. Another inherent feature of this vectorizer was that it performed automatic tokenization, lowercase conversion, and punctuation handling, thus providing basic text pre-processing functionality. In addition to this we also converted numbers to a generic number token, this to reduce the variability in the text and on the assumption that the actual numeric values had limited value for our classification task. Stemming was considered but since we wanted to preserve verb tenses we decided against this. An example of a case where verb tense could make a difference would be the discussion of a planned CVC insertion versus an actual insertion. For this reason we set up a separate experiment to investigate the effect of stemming. Handling of negation is another common challenge in natural language processing tasks. We did not make any efforts towards explicitly handling negation, assuming instead that the use of n-gram models would enable the classifiers to differentiate between negated and non-negated concepts. As for n-gram models, we experimented with different n-gram dimensions and their impacts on classifier performance. In the end we settled on using 1- to 3-grams for all experiments as this combination seemed to provide the best results. The use of unigrams was partly motivated by the terseness of clinical language; single-word features could make a difference as single-word sentences were known to exist.

We selected a set of common algorithm implementations in scikit-learn using the default or recommended settings as the initial parameters. For the first experiment we wanted to see how the number of features used would affect the performance of the selected algorithms on the majority class *Use* and the minority class *Rem*. A key aspect when limiting the number of features used is how features are selected. We decided to use the scikit-learn `SelectKBest` univariate feature selector with a chi-squared statistical test for scoring. This selector scores the features according to the chosen scoring function and returns the desired number of features. Manual inspection of the top features showed that the chosen features made sense given the context and our domain knowledge. For example, direct references to CVC or various catheter types, were highly ranked. Also, many features were closely associated with nursing tasks, e.g. the removal of sutures. This was reasonable since we had a large number of nursing notes in our data set.

Table 7 shows an example of the 20 highest ranked features, translated from Norwegian to English, in a trial experiment on all 34,810 sentences.

Table 7: Highest scoring features

cvc	cvc day	cvc care
removed sutures	removed sutures from	from hickman catheter
given cvc	cvc was inserted	received new cvc
have been inserted	hickman	hickman catheter
new cvc	disc cvc	discontinuing cvc
discontinued cvc day	discontinued cvc	care
sutures from	sutures from hickman	

In Figure 5 we see the balanced F_1 score for the *Use* class for the chosen algorithms while Figure 6 shows the same experiment for the sparse *Rem* class. For both classes the performance of the `linear_svc_11`, `linear_svc_12` and `ridge` algorithms improves with the number of features. For the *Rem* class there are better performing algorithms that seem to perform well with a limited number of features. In terms of priorities, we decided that optimizing performance for the *Use* class should be our primary experiment objective. Having a well-performing CVC usage classifier would not only be beneficial for the purpose of counting days of CVC use but would also make good use of the more prevalent *Use*-related annotations in our data set. For these reasons we opted to use the `linear_svc_11` and `linear_svc_12` algorithms for the remainder of our experiments. These algorithms are the scikit-learn implementations of a linear kernel support vector machine (SVM) with parameters `loss=squared_hinge`, `penalty=11` (or `12`), `dual=False` and `tol=1e-3`. The `11` and `12` influences the sparsity of the internal coefficient vectors.

We repeated the number of features experiment although this time only using the `linear_svc_11` algorithm. Figure 7 shows the F_1 -score for all 5 classes. As could be expected, the prediction performance is lower for the classes with less training data. Coincidentally the predictive quality of the *Use* classifier is similar to the results seen in the adverse CVC event detection by Penz et al. (2007), although a direct comparison can not be made.

The next experiment sought to evaluate if including sentence section information (see Figures 3 and 4) and note type as features could improve classifier performance. The simplest way to achieve this was to use the scikit-learn `FeatureUnion` functionality which combines different feature sources into a unified feature vector. Applying this on each sentence would give us a combined feature vector that relied on both the standard bag-of-words features as well as additional section and note type features. We chose to give each

feature source equal weight rather than weighting some of them as more important than others. We defined three experiment setups with different feature source combinations: sentence, sentence + section, and sentence + section + note type. Table 8 shows the results, where F_1 , precision, and recall are given for each setup.

Table 8: Experiments combining sentence and note type information

Class	Sen			Sen/Sec			Sen/Sec/Not		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
None	99.8	99.9	99.8	99.8	99.8	99.8	99.8	99.8	99.8
Plan	63.4	38.2	47.5	69.2	41.1	51.2	66.9	38.1	48.4
Ins	47.5	18.1	24.6	50.0	18.1	25.5	50.0	19.8	27.5
Use	74.0	84.5	78.9	74.3	85.4	79.5	73.9	85.0	79.1
Rem	81.3	22.4	35.0	81.3	26.6	39.1	81.3	22.4	35.0

The numbers shown in bold are the highest scores for the given class. Most noteworthy is that adding section features has a positive effect on prediction quality while note type has a negative effect. The one exception for the latter is the *Ins* (insertion) class. A manual inspection revealed that documentation of CVC insertion was almost always found in the anesthesiology record note type, so in that way it made sense that including note type information would have a positive impact.

In our final experiment we wanted to investigate the effect of stemming on classifier performance. The effect of pre-processing techniques such as stemming may be highly dependent on e.g. the text domain and the language used (Uysal and Gunal, 2014). Using `linear_svc_12` we ran an experiment with and without stemming, otherwise using all available text features, 4-fold cross-validation and no sentence or note type features. For stemming we used the Norwegian Snowball stemmer that is bundled with NLTK. The results can be seen in Table 9.

Table 9: Experiments with and without stemming

For the most common class, *Use*, stemming has a slightly negative although negligible effect. For the sparsest class, *Rem*, there is however a

Class	Without stemming			With stemming		
	Pr	Re	F1	Pr	Re	F1
None	99.7	99.9	99.8	99.7	99.9	99.8
Plan	80.0	46.1	57.9	82.9	48.7	60.7
Ins	48.1	34.3	38.5	48.1	34.3	38.5
Use	77.5	78.5	78.0	77.5	77.9	77.7
Rem	63.2	31.8	41.7	66.5	38.4	48.5

marked improvement with stemming. A similar effect is seen with *Plan*, which is also quite sparse. For *Ins* there is no difference. A possible explanation is that a potential benefit from differentiating between e.g. verb tenses is outweighed by the dimensionality reduction of the feature space that stemming provides in our quite small data set. A similar effect for another non-English language is seen in e.g. Torunoglu et al. (2011), where stemming was found to be beneficial for small training sets.

7 Conclusion

We found that even with limited training data it is still possible to predict CVC use events from sentences in clinical notes with adequate precision and recall. This gives us a foundation for later inference of CVC usage periods, thus allowing us to get better estimates for the number of days that CVC has been in use. It seems likely that additional training data will improve classifier performance; in particular, it would be useful with better performance for insertion and removal events. Another interesting finding was that using sentence context information would provide an additional performance boost. It can reasonably be assumed that more accurate classifiers for the sparser events will lead to less ambiguity when attempting to map CVC use intervals from discrete CVC events.

There are several avenues for further research on this topic, most importantly the aforementioned CVC usage period prediction and day count. In addition there are many possible approaches towards strengthening the event prediction foundation. Given the sparsity of training data, one interesting option would be to see how convolutional neural networks perform, given that they have been shown to sometimes work well even with limited training data sets. Another option is to expand our notion of sentence context to also include data from previous clinical notes, thus providing even more background that may aid the classifiers. The key here is probably to find a representation of previous events, treatment and patient background that is

both at a high-enough level to be useful but also not overly simplistic.

Another observation is on the difficulty of extracting text from EHR systems and the fact that exporting options are often quite limited. While the importance of secondary use of clinical data is increasingly recognized (Meystre et al., 2017), there are often many practical obstacles towards accessing such data. The trend among EHR vendors is somewhat towards e.g. interoperability and API access, but often only for structured content. For research on unstructured clinical text, it is also necessary that the text is available in a format useful for export and that elements of text structure and the usage context are not lost during the export process. In particular, text as part of forms lose their meaning unless the specific form is also available for text processing.

The end goal of this project is as mentioned to improve our knowledge of the prevalence and duration of CVC use in hospitals. The work described in this paper is preliminary and can be considered a means towards this end. A key element is to be able to accurately identify transitions between CVC use states: from planning to insertion, care during use, and removal. When doing so it is important to recognize the difference between the actions that were originally applied to the patient and how these were ultimately documented. There are multiple aspects that must be taken into account, not least given the variety of note types. For example, a nursing note will typically describe actions and observations from the current 8-hour shift and which are relevant for the next shift. These notes are mostly descriptive, and will also be written shortly after the described events took place. On the other hand, a discharge note will summarize a wider variety of events that took place over a longer period of time. It may also be more reflective and also outline plans for further treatment. Mapping descriptions in the clinical notes as accurately as possible to the points on the timeline where they actually took or will take place is critical for getting an accurate CVC use day count.

8 Acknowledgements

The reported research was approved by the regional ethical committee (Reference 2010/338) and supported by EVICARE (Evidence-based care processes: Integrating knowledge in clinical information systems) (NRC project 193022). To a lesser extent, the BigMed project (NRC project 259055) also contributed. This article is an extended version of an article accepted at BIBM 2018 Røst et al. (2018).

9 Figures

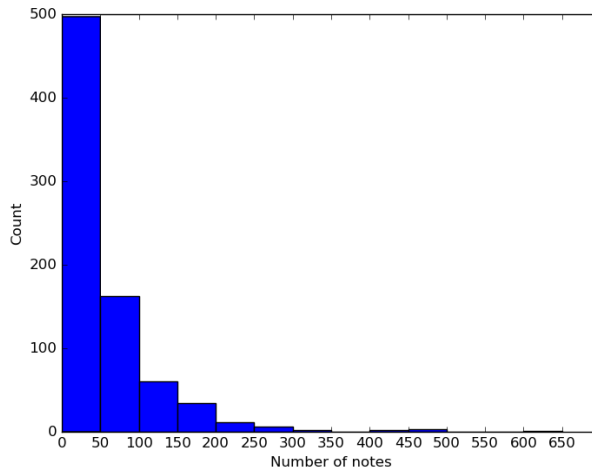


Figure 1: Episode of care length (notes)

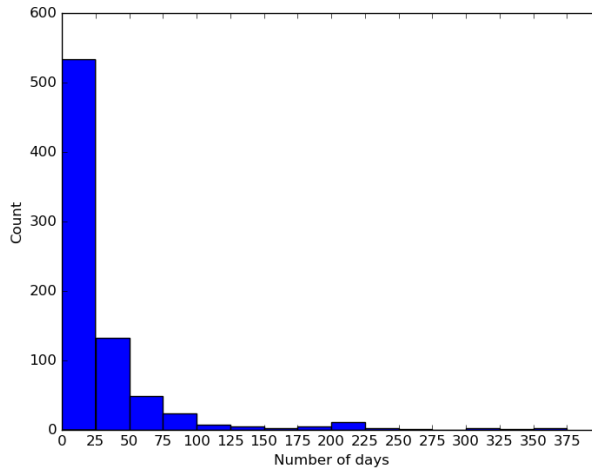


Figure 2: Episode of care length (days)

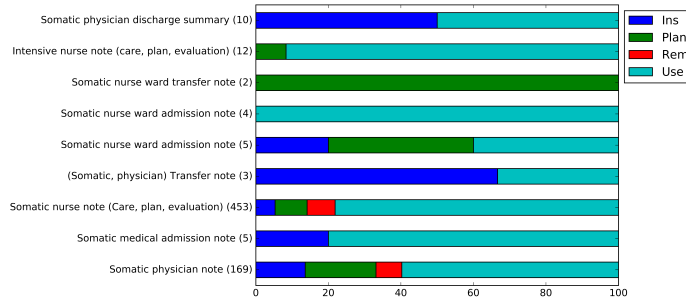


Figure 3: Aggregate class use per document type

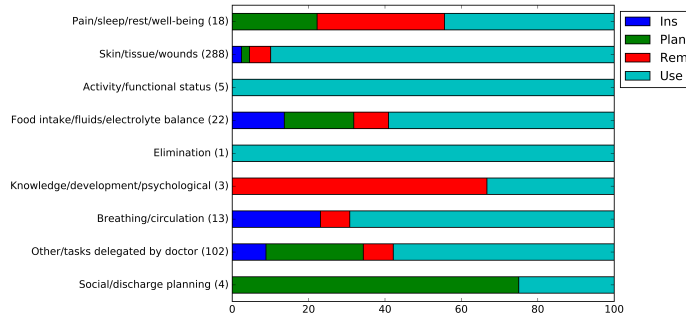


Figure 4: Aggregate class use per nursing note section type

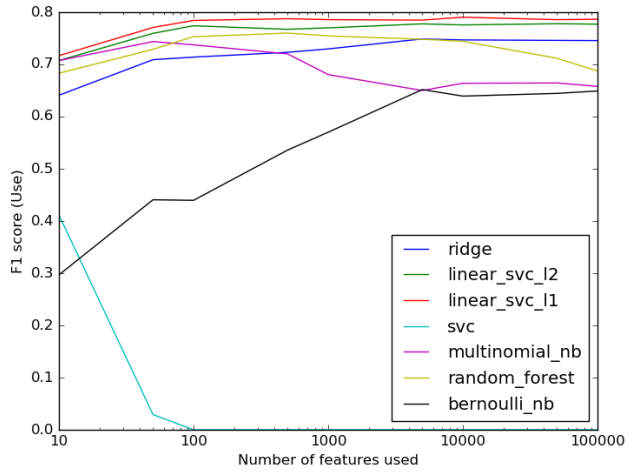


Figure 5: F1 vs. number of features (Use)

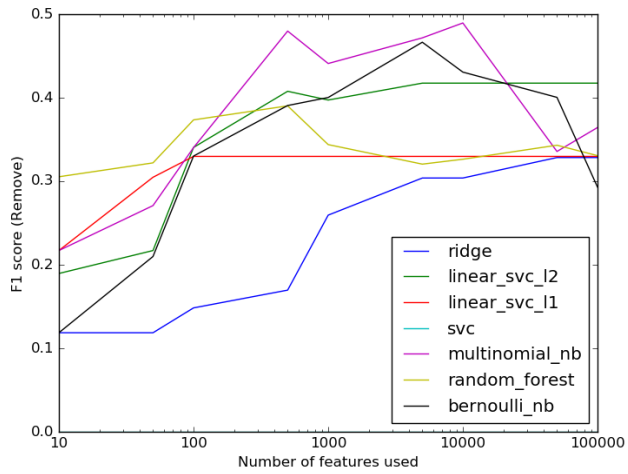


Figure 6: F1 vs. number of features (Rem)

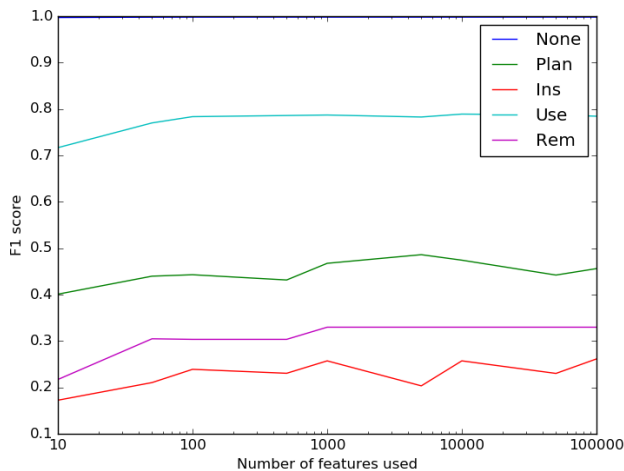


Figure 7: F1 vs. number of features (all classes)

10 Bibliography

References

- (2008). Lov 20. juni 2008 nr. 44 om medisinsk og helsefaglig forskning.
- Bates, D. W., Evans, R. S., Murff, H., Stetson, P. D., Pizziferri, L., and Hripcsak, G. (2003). Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115–128.
- Berg, I. A. (2014). Automated annotation of events related to central venous catheterization in norwegian clinical notes. Master’s thesis, Norwegian University of Science and Technology.
- Bruin, J. S. D., Blacky, A., and Adlassnig-Peter, K.-P. (2012). Assessing the clinical uses of fuzzy detection results in the automated detection of cvc-related infections: a preliminary report. *Studies in Health Technology and Informatics*, 180(Quality of Life through Quality of Information):579–583.

- Brun-Buisson, C. (2001). New technologies and infection control practices to prevent intravascular catheter-related infections. *American Journal of Respiratory and Critical Care Medicine*, 164(9):1557–1558.
- Cohen, E. R., Feinglass, J., Barsuk, J. H., Barnard, C., O'Donnell, A., McGaghie, W. C., and Wayne, D. B. (2010). Cost savings from reduced catheter-related bloodstream infection after simulation-based education for residents in a medical intensive care unit. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 5(2):98–102.
- Fletcher, S. (2005). Catheter-related bloodstream infection. *Continuing Education in Anaesthesia Critical Care & Pain*, 5(2):49–51.
- Govindan, M., Citters, A. D. V., Nelson, E. C., Kelly-Cummings, J., and Suresh, G. (2010). Automated detection of harm in healthcare with information technology: a systematic review. *BMJ Quality & Safety*, 19(5):e11–e11.
- Hojsak, I., Strizić, H., Mišak, Z., Rimac, I., Bukovina, G., Prlić, H., and Kolaček, S. (2012). Central venous catheter related sepsis in children on parenteral nutrition: a 21-year single-center experience. *Clinical Nutrition*, 31(5):672–675.
- Husby, H. (2014). Klassifisering av sykepleiejournalen - kan kunnskap om sykepleiedokumenter forbedre gjenkjenning av hendelser knyttet til sentralvenekateterisering? Master's thesis, Norwegian University of Science and Technology.
- Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.
- Loper, E. and Bird, S. (2002). Nltk. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*.
- Løwer, H. L., Eriksen, H.-M., Aavitsland, P., and Skjeldestad, F. E. (2013). Methodology of the norwegian surveillance system for healthcare-associated infections: the value of a mandatory system, automated data collection, and active postdischarge surveillance. *American Journal of Infection Control*, 41(7):591–596.
- Maki, D. G., Kluger, D. M., and Crnich, C. J. (2006). The risk of bloodstream infection in adults with different intravascular devices: a systematic review

- of 200 published prospective studies. *Mayo Clinic Proceedings*, 81(9):1159–1171.
- McKibben, L., Horan, T., Tokars, J. I., Fowler, G., Cardo, D. M., Pearson, M. L., and Brennan, P. J. (2005). Guidance on public reporting of healthcare-associated infections: Recommendations of the healthcare infection control practices advisory committee. *American Journal of Infection Control*, 33(4):217–226.
- Mermel, L. A. (2017). Short-term peripheral venous catheter-related bloodstream infections: a systematic review. *Clinical Infectious Diseases*, 65(10):1757–1762.
- 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.
- Michelson, J. D., Pariseau, J. S., and Paganelli, W. C. (2014). Assessing surgical site infection risk factors using electronic medical records and text mining. *American Journal of Infection Control*, 42(3):333–336.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2012). Scikit-learn: Machine learning in python. *CoRR*.
- Penz, J. F., Wilcox, A. B., and Hurdle, J. F. (2007). Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182.
- Pronovost, P. J., Goeschel, C. A., Colantuoni, E., Watson, S., Lubomski, L. H., Berenholtz, S. M., Thompson, D. A., Sinopoli, D. J., Cosgrove, S., Sexton, J. B., Marsteller, J. A., Hyzy, R. C., Welsh, R., Posa, P., Schumacher, K., and Needham, D. (2010). Sustaining reductions in catheter related bloodstream infections in michigan intensive care units: Observational study. *BMJ*, 340(feb04 1):c309–c309.
- Røst, T. B., Tvedt, C. R., Husby, H., Berg, I. A., and Nytrø, Ø. (2018). Capturing central venous catheterization events in health record texts. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 488–495.

- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsuji, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.
- Taylor, R. W. and Palagiri, A. V. (2007). Central venous catheterization. *Critical Care Medicine*, 35(5):1390–1396.
- Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S., and Gurbuz, M. Z. (2011). Analysis of preprocessing methods on classification of turkish texts. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 112–117.
- Trick, W. E., Chapman, W. W., Wisniewski, M. F., Peterson, B. J., Solomon, S. L., and Weinstein, R. A. (2003). Electronic interpretation of chest radiograph reports to detect central venous catheters. *Infect Control Hosp Epidemiol*, 24(12):950–954.
- Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.
- Wong, A. V., Arora, N., Olusanya, O., Sharif, B., Lundin, R. M., Dhadda, A., Clarke, S., Siviter, R., Argent, M., Denton, G., Dennis, A., Day, A., Szakmany, T., and group, T. F. I. C. N. A. P. I.-. (2018). Insertion rates and complications of central lines in the uk population: A pilot study. *Journal of the Intensive Care Society*, 19(1):19–25.

PAPER C

*Using Neural Networks to Support High-Quality
Evidence Mapping*

RESEARCH

Open Access

Using neural networks to support high-quality evidence mapping

Thomas B. Røst^{1*} , Laura Slaughter¹, Øystein Nytrø¹, Ashley E. Muller² and Gunn E. Vist²

From 14th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO 2020) Virtual. 19 October 2020

*Correspondence:

brox@ntnu.no

¹Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Full list of author information is available at the end of the article

Abstract

Background: The Living Evidence Map Project at the Norwegian Institute of Public Health (NIPH) gives an updated overview of research results and publications. As part of NIPH's mandate to inform evidence-based infection prevention, control and treatment, a large group of experts are continuously monitoring, assessing, coding and summarising new COVID-19 publications. Screening tools, coding practice and workflow are incrementally improved, but remain largely manual.

Results: This paper describes how deep learning methods have been employed to learn classification and coding from the steadily growing NIPH COVID-19 dashboard data, so as to aid manual classification, screening and preprocessing of the rapidly growing influx of new papers on the subject. Our main objective is to make manual screening scalable through semi-automation, while ensuring high-quality Evidence Map content.

Conclusions: We report early results on classifying publication topic and type from titles and abstracts, showing that even simple neural network architectures and text representations can yield acceptable performance.

Keywords: Evidence maps, Evidence based medicine, Knowledge dissemination, Automated coding, Machine learning, Deep learning

Background

Experts, policy makers and researchers worldwide are scrambling to keep up with the influx of potentially relevant COVID-19 studies. Research is being published at an unprecedented pace and in volumes never seen before. Whereas a traditional peer review- and journal-based publication process would take 6–12 months, research findings now often find their way to readers in a matter of days or weeks. The use of preprint servers, with only cursory quality checks, is increasing. While this has had a positive impact on knowledge dissemination speed in the medical sciences, this arguably comes at a cost to quality, reliability and trustworthiness [1].



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The need for timely, informed and quality-assessed knowledge is widely recognised as crucial for handling the ongoing COVID-19 pandemic. One initiative to meet this need, known as the Living Evidence Map Project, was launched at the Norwegian Institute of Public Health (NIPH) within their Division for Health Services [2]. The NIPH has a large team of experienced review authors that regularly conducts systematic reviews of medical science research, this as part of their mandate to inform evidence-based decisions pertaining to prevention and infection control. Evidence maps provide a useful overview of the literature, but since many of the workflow steps overlap, they can be seen as a precursor towards the production of systematic reviews.

The challenge with the current approach to evidence mapping is that it is currently mostly manual and requires considerable amounts of expertise from the reviewers. This leads to a review and coding process that is already struggling to keep up with the volume of incoming publications and that is hard and costly to scale. We believe that technologies of medical language processing, knowledge extraction and machine learning have the potential to assist and amplify the expertise required to produce systematic reviews and evidence maps.

Automation was not introduced to synthesising medical evidence in the past since it was thought to be inadequate and would potentially only increase the amount of work needed, adding effort and time to check over machine results. We began with exploring past work that assessed the use of text mining to support systematic review workflows. Projects from years prior to COVID focused on the literature screening phase of the work process and some have been implemented in the current reviewing support systems [3].

Although the screening tools have been implemented into workflows, NIPH has no automated support that would speed up coding procedures. We have initiated a series of experiments to explore multi-label deep learning classification to help with this task. In this work, we conduct four experiments in order to assess the possibilities of using deep learning techniques in the evidence mapping workflow. Automated approaches are evaluated based on measurements of precision and recall, however, we know very little about what this means to those who wish to implement automation into workflows where a high-quality knowledge product is expected as the result. We hope to learn more about acceptable error rates when applied to a real-life needs and workflows. Therefore, our research is centred around the question of whether it is possible to reduce manual efforts while at the same time maintaining high-quality evidence maps.

We received a set of training data from NIPH as a result of their manual coding to produce evidence maps for COVID-19, and we focused our research questions on exploring the classification of publications for automated coding. Our main research question centres on performance when using deep learning models to classify the COVID-19 research literature: *Can we expect an accurate classification of clinical research topics, publication type, and data types using only publication titles and abstracts?*

We believe that this work will provide the groundwork for understanding the implementation of machine learning and deep learning techniques in real-life clinical scenarios and workflows. The Allen Institute for Artificial Intelligence put forth the COVID-19 dataset [4] which is a set of scientific publications available related to COVID-19 as well as related historical coronavirus research, including SARS and MERS. We make use

of this dataset in our experiments, and seek to tie the challenges outlined to the needs related to evidence mapping and generation of systematic reviews.

The NIPH coding workflow

In the NIPH coding workflow, incoming articles that will be coded for inclusion in the evidence maps have been screened and quality controlled. In this context the term “coding” means manual classification of each article according to a predefined set of discrete categories, e.g. paper topic (diagnosis, etiology etc.) or data type (primary data, secondary data etc.), some of which will be discussed in more detail later. The articles include those that their collaborator, EPPI-Centre [5], has already screened (using a combination of machine learning and manual methods.) These are supplemented with studies from NIPH’s own searches. As of 15 July 2020, the map contains 6513 publications categorised by topic, population, and publication type.

The categorisation process is labour-intensive. Depending on the study, it may take 3–15 min to code a paper. Studies from the corpus are randomly allocated to two coders, a “core coder” and an “external coder.” The core coder has the ability to see the external coder’s coding. When the external coder codes first, the core coder can see those codes while coding themselves. The process is manual and is done to increase the speed of work for the core coder who has greater expertise.

The breadth of studies brings with it many publication types and topics that are not always easy to categorize. In the case of disagreement between coders, differences are discussed in a reconciliation meeting. This process usually means that the core coder’s codes are adopted as the final version, and sometimes with some extra codes added after input from the external. This is on-the-ground learning for the external coder, because it’s really the only time they see how a study “should” be coded. The benefit of this process is a continuous overview over the consistency in the coding efforts. Differences in coding may be due to different viewpoints or human error, but were mostly resolved without the need of a third researcher to adjudicate.

NIPH has 28 coders in total, all with a research or medical background. There are 15 external volunteers and 13 from the Norwegian Institute of Public Health. There is a programme to train new coders with the coding process, with an introduction to the necessary software and the NIPH coding manual. Following the training, new coders are then paired up with an experienced coder to continue training on-the-job, with weekly discussions for general questions or specific studies. During these discussion rounds, any disagreements in coding are discussed and reconciled. Currently coders are managing to code over 10 studies per hour.

NIPH has created its own coding system with an accompanying manual describing all codes in detail, this with the aim of reducing ambiguity. The NIPH coding manual has developed throughout the project to address the developing research. This dynamic approach has allowed for much-needed flexibility, but at times this can require considerable work to realign older codes.

Related work

Deep learning machine learning models are seeing increased use for a wide variety of natural language processing (NLP) tasks [6], motivated by the ability to produce results

that improve on the performance of previous-generation machine learning methods. For text classification, they have surpassed traditional methods for tasks such as sentiment analysis, news categorization and natural language inference [7]. In the medical domain, there has been a great deal of focus on deep learning for medical image processing [8] but other avenues of research are continuously opening up. For example, it has been shown to perform well for identifying relevant publications from medical literature, especially when considering that less time is spent on e.g. feature engineering and MeSH term linking [9]. Convolutional neural networks, a particular deep learning architecture, have shown promising performance for tasks such as automated ICD-9 coding [10] and de-identification of clinical texts [11]. A recent study concluded that the use of deep learning methods has yet to fully penetrate clinical natural language processing but also that usage was increasing rapidly [12].

Production of evidence maps, systematic reviews as well as best practice guidelines have been discussed in terms of the ecosystem of healthcare system data. Connecting and reusing health data is an essential aspect to implementing precision medicine. The flow of data from patient care and clinical trials to published results and observations, and through the cycle of summarization and reuse to inform care has also been connected to the concept of learning healthcare systems. Even prior to the COVID-19 crisis, the issues and problems have been identified as evolving and cutting-edge research [13]. Recently published discussions on evidence ecosystems call for more coordinated and integrated synthesis that is relevant, trustworthy, and useful for decision making [14].

Recent publications and work from 2016 until the present time has originated from the group at the National Centre for Text Mining, University of Manchester. They focused their research on the literature screening phase of the systematic review process, with methods developed for prioritizing references [15], document clustering using a predictive network [16], and topic detection [17]. In addition, they built a prototype based on the sum of their work, Robot Analyst. The work was completed as part of a funded project titled Supporting Evidence-based Public Health Interventions using Text Mining with collaboration of the University of Liverpool Machine Learning and Data Analytics group, and the National Institute for Health and Care Excellence (NICE). This group was influential and together with the EPPI systematic review tool developers at UCL, EPPI implemented screening functionality into their production system.

There are several other examples of research on reducing the manual effort associated with classification of scientific literature. A 2006 study by Cohen et al. used machine learning for automated classification of document citations, this with the purpose of aiding experts in updating system reviews of drug class efficacy [18]. Moving beyond the medical domain, work has been done on e.g. classification of mathematical research [19] and on general research literature with the purpose of applying the correct Dewey Decimal Classification code [20]. While much work focuses on classification of English-language literature, examples of using machine learning methods for automated coding of scientific literature in the Russian language [21]. Most approaches appear to be based on supervised learning but use of unsupervised learning also exists [20].

Methods

The coding data was exported as a JSON file on May 4th, 2020, from the EPPI-Mapper [22] tool used by NIPH. It had two main sections, `CodeSets` and `References`, containing respectively the coding definitions and the publications with the applied codes. To get a better feel for the type and volume of data available to us we analyzed the data coverage in the `CodeSets` and `References` sections.

The codes in the `CodeSets` section has a tree structure where each attribute node has an ID, a name, a description, a set ID, a set description and a type. Parent nodes also has a list of child attributes. In total there are 40 parent attributes and 223 leaf attributes.

There was a total of 1332 references. Each reference section had a number of metadata fields. Some of these, such as *Abstract*, *Authors* and *Volume* were directly related to the associated publication. The rest of the fields, such as *Codes* and *Comments* contained data that had been added by the coders during the publication coding process. Table 1 shows an overview of how well these fields are covered in the dataset. We see that *Codes* coverage is complete, as could be expected, while *Comments* and *Keywords* are more sparsely used. As for bibliographic data, the publication *Title* is always present while some *Abstract* entries are missing. The journal or conference title is found in the *Parent-Title* field which is absent in only 1 case. The *Itemid* field was confirmed to be unique. The *Keywords* value, if available, contains a newline separated list of coder-provided keywords. In conversations with NIPH we learned that their coders did not add keywords and that the origin of this information is therefore of uncertain quality.

For 181 reference entries the abstract is missing, which means that very little textual data beyond the title, keywords and comments is available as classification features. To alleviate this we attempted to link the references without abstracts to the COVID-19 Open Research Dataset (CORD-19) [4], specifically to the main `metadata.csv` file, using publication title and DOI as linking identifiers. The results can be seen in Table 2, showing that matching on DOI performed the best but even then only 33 missing abstracts could be found. We did not do any normalization on the link values apart from converting to lower case so it is possible that some links were missed this way. It is also possible that in many cases abstracts are simply not available. We decided to augment the data used for the experiments with the additional abstracts found from DOI matching so as to maximize the amount of data available to us.

To get a feel for how the coding practice has evolved we looked closer at the *DateCreated* (when a reference was imported into the system) and *DateEdited* (when a reference was coded) fields. Figure 1 shows how many references were imported and coded per week for the duration of the data sets. The number of coded references increase towards the latter half of the period. The team started with 4 coders and by week 14 this number had increased to 10, the majority of whom were part-time coders. Note that using this field to assess initial coding time is not entirely accurate as sometimes the post-coding quality control process would lead to a reference being recoded. There is also a chance that any administrative changes to other fields could impact the value of the *DateEdited* field.

We then looked at how the actual codes were distributed across the code hierarchy. Many of the codes had an associated *AdditionalText* field with comments made by the coder; coders were instructed to use this field to flag things for discussion during the

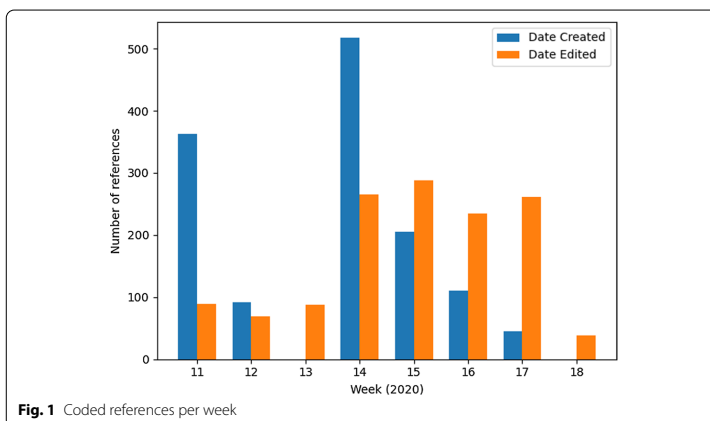
Table 1 Reference field coverage

Key	# References	Cov. (%)
Abstract	1151	86.41
Authors	1321	99.17
Availability	0	0.00
City	9	0.68
Codes	1332	100.00
Comments	689	51.73
Country	0	0.00
CreatedBy	1324	99.40
DOI	1216	91.29
DateCreated	1332	100.00
DateEdited	1332	100.00
EditedBy	1332	100.00
Edition	0	0.00
Institution	464	34.83
Issue	316	23.72
ItemId	1332	100.00
ItemStatus	1332	100.00
ItemStatusTooltip	1332	100.00
Keywords	446	33.48
Month	1	0.08
OldItemId	1332	100.00
Outcomes	0	0.00
Pages	513	38.51
ParentAuthors	0	0.00
ParentTitle	1331	99.92
Publisher	8	0.60
ShortTitle	1332	100.00
StandardNumber	506	37.99
Title	1332	100.00
TypeName	1332	100.00
URL	943	70.80
Volume	449	33.71
Year	1332	100.00

Table 2 Matching FHI data with COVID-19 data

Match element	Matches	Matches w/ abstr.
Title	123	28
DOI	127	33

coding reconciliation process. Other than that there was no additional metadata. Table 3 shows some key numbers regarding code use. Several of the code attribute IDs could not be found in the code set. For some reason some of the coded attribute IDs mapped to the attribute set IDs rather than the standard attribute IDs; this needs to be investigated further.

**Table 3** Coding statistics

Number of codes	25,133
Avg. number of codes per ref.	18.9
Number of comments	265
Number of unknown codes	1234
Number of unique unknown codes	25

Table 4 Root code use

Class	# Refs
Publication type, detailed	1332
Publication type	1332
Data type	1332
Population	1332
Topic	1332
Topic: Diagnosis	454
Topic: Aetiology	452
Topic: Prognosis	424
Topic: Prevalence	206
Topic: Interventions to treat the infected patient	162
Topic: Interventions targeted at system level to improve management of the pandemic	143
Topic: Experiences and perceptions; consequences; social, political, economic aspects	126
Topic: Infection prevention and control	119

For each code found in a reference we stored the reference ID so as to know how many references are available for a given code. Table 4 shows an overview of the top-level codes and for how many references in the full data set these codes have been applied, sorted by the number of coded references. Codes used only once or never are not shown. We see that data coverage is incomplete for all but the top 5 top-level codes. The more

specific topic codes in the bottom half of the table contained more detailed drilldown into the various subtopics. In our experiments we ended up focusing on the *Topic*, *Data Type* and *Population* codes, the main reasons being that they all had single-level coding hierarchies and reasonably well distributed classes.

Results

For our experiments we made a selection of codes that we thought would be best suited as classification labels, with the selection criteria being the amount of data and overall class balance. All experiments were run using the Keras deep learning framework [23]. Keras was chosen because it is a popular and increasingly used framework which also provides several convenience functions to lessen the workload both for text pre-processing and the general machine learning experiment workflow. We relied on the default TensorFlow [24] symbolic math library backend. The choice of deep learning machine learning methods is not only motivated by recent performance advances but also because they usually reduce the need for activities such as feature engineering [25]. For the rest of the discussion, we will refer to codes as labels given that this is a more common terminology for classification tasks.

Experiment 1: classifying topics from the publication title

For our initial experiment we wanted to build a classifier for the *Topic* label, as shown in Table 5. The goal is to correctly classify the topic based on information available to us, such as the publication title, the abstract, the publication outlet and so on. The total number of applied labels is 2084. Since this exceeds the number of references it follows that some of the references must be labelled with multiple topics, making this a multi-label classification task. While some topics occur more often than others, there are no topics that are exceptionally scarce and the dataset is relatively balanced.

We first attempted to use only the publication title as input for our features. The primary benefit of using the title is that we know that it always will be present in the dataset. We put all titles into the Keras `Tokenizer` API, which splits on whitespace, removes punctuation, lowercases all tokens and outputs a bag-of-words-encoded feature matrix with a selected output mode. Each row in the matrix has a vector with the size of the vocabulary, with each word having its own position. In our case we relied on the `count` mode which means that the word frequency is used as a feature value.

Table 5 Reference count (Topic)

Class	# Refs
Topic	1332
Prevalence and incidence	205
Etiology	452
Diagnosis	454
Infection prevention and control	119
Interventions to treat the infected patient	162
Interventions targeted at system level	142
Prognosis	424
Experiences and perceptions; consequences; social, political, economic aspects	126

For all experiments we set a maximum vocabulary size of the 1,000 most frequent words in the corpus.

We decided to start with the simplest possible neural network architecture, using a sequential model with three dense layers, each having 64 units. The number of layers was set after some initial experimentation to make sure that the model would not be lacking in representational power. The number of layers and units was motivated by similar classification examples as described by the creator of Keras [25] rather than previous experience. Each layer used the `relu` activation function. The fourth and final layer had 8 units, corresponding to the number of classes, and a `sigmoid` activation function. As per recent best practices for this type of classification task we used the Adam optimization algorithm, a `binary_crossentropy` loss function and, since this was a multi-label classification problem, the `categorical_accuracy` evaluation metric. We also added precision and recall metrics as these would be more useful in practice for evaluation purposes. Batch size was set to 128. Since we had a fairly small amount of data to work with we used 4-fold cross validation for all experiments, averaging the results. The number of epochs was set by doing trial runs with 20% of the training data set aside for validation. We then observed the loss function output and chose the final number of epochs to roughly correspond with the loss function minimum, this to avoid overfitting. For the final run we used all available data for training and ignored validation. When evaluating on the data set aside for testing we would end up with a vector of values between 0.0 and 1.0. If the value was above a threshold of 0.5 we interpreted this as a positive prediction for the given class.

Table 6 shows the classification results for the *Topic* label in the form of average precision, recall and F1 metrics as well as the standard deviation. We see that precision is in general better than recall, while recall seems to be positively correlated with the amount of training data. For the classes with less data the precision standard deviation is high and results would fluctuate considerably between each run. The difference between precision and recall performance could be explained by the lack of data in the titles: the model picks up on commonly occurring words which makes for safe predictions while the majority of titles have too little information to make a good prediction.

Table 6 Topic classification results from title

Class	Precision	Recall	F1
Diagnosis	0.71 (0.06)	0.61 (0.08)	0.66 (0.07)
Etiology	0.69 (0.04)	0.52 (0.08)	0.59 (0.04)
Experiences and perceptions; consequences; social, political, economic aspects	0.80 (0.10)	0.38 (0.07)	0.51 (0.07)
Infection prevention and control	0.72 (0.23)	0.11 (0.02)	0.18 (0.03)
Interventions targeted at system level	0.23 (0.27)	0.08 (0.12)	0.11 (0.17)
Interventions to treat the infected patient	0.73 (0.06)	0.39 (0.04)	0.51 (0.04)
Prevalence and incidence	0.63 (0.07)	0.28 (0.03)	0.39 (0.04)
Prognosis	0.73 (0.05)	0.61 (0.04)	0.66 (0.02)

Table 7 Topic classification results from abstract

Class	Precision	Recall	F1
Diagnosis	0.72 (0.09)	0.68 (0.06)	0.70 (0.03)
Etiology	0.69 (0.06)	0.50 (0.04)	0.58 (0.04)
Experiences and perceptions; consequences; social, political, economic aspects	0.79 (0.08)	0.45 (0.05)	0.57 (0.06)
Infection prevention and control	0.77 (0.13)	0.21 (0.05)	0.33 (0.07)
Interventions targeted at system level	0.65 (0.21)	0.15 (0.07)	0.23 (0.10)
Interventions to treat the infected patient	0.75 (0.08)	0.38 (0.04)	0.50 (0.04)
Prevalence and incidence	0.71 (0.06)	0.30 (0.06)	0.42 (0.05)
Prognosis	0.74 (0.02)	0.56 (0.06)	0.63 (0.03)

Table 8 Reference count (Publication type)

Class	# Refs
Publication type	1332
Systematic reviews	156
Studies and modelling	1051
Non-systematic reviews and others	194

Experiment 2: classifying topics from the publication abstract

For the next experiment we stuck with the *Topic* coding from experiment 1 but switched the input data source from the title to the abstract. This would presumably give the deep learning model more data to work with. We kept all other tokenization parameters and model hyperparameters equal, including the network architecture.

Results from classifying the *Topic* label based on abstracts are found in Table 7. The most noticeable difference is that both precision and recall have improved for the classes that performed poorly in the previous experiment. The abstract will in most cases be substantially longer than the title and as such there is more information to work with for the neural network model, thus improved performance is as expected.

Experiment 3: classifying publication type from the publication abstract

We repeated the same experiment, using the abstracts as a basis for our features but this time attempting to classify for the *Publication type* label. The class distribution can be seen in Table 8. As before, this is a multi-label classification task but this time publications are much more likely to have a single label applied.

With all parameters from the previous experiments kept equal the results are shown in Table 9. Performance for the *Studies and modelling* class is best but this is also by far the most prevalent class.

Experiment 4: classifying data type from the publication abstract

This experiment was again similar to the previous one but now for the *Data type* label. This label says something about the kind of data, if any, that was used in the publication. Table 10 shows the class distribution and that most of the reviewed publications deal with primary data.

Results of this experiment can be seen in Table 11.

Table 9 Publication type classification results from abstract

Class	Precision	Recall	F1
Non-systematic reviews and others	0.52 (0.41)	0.05 (0.06)	0.08 (0.11)
Studies and modelling	0.86 (0.02)	0.98 (0.01)	0.92 (0.01)
Systematic reviews	0.91 (0.07)	0.53 (0.08)	0.67 (0.07)

Table 10 Reference count (Data type)

Class	# Refs
Data type	1332
Primary data	789
Secondary data	231
Modelled/computed	271
No data (i.e. comment, editorial)	74

Table 11 Data type classification results from abstract

Class	Precision	Recall	F1
Modelled/computed	0.77 (0.09)	0.66 (0.05)	0.71 (0.04)
No data (i.e. comment, editorial)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Primary data	0.80 (0.02)	0.92 (0.02)	0.86 (0.01)
Secondary data	0.89 (0.04)	0.50 (0.01)	0.64 (0.01)

Experiment 5: classifying topics from the publication abstract with CNN and pre-trained word embeddings

The final experiment is a repeat of experiment 2 but this time with a more advanced architecture which is also supported by pre-trained word embeddings. We used the 100-dimensional GloVe embeddings of 400K words which is based on data from Wikipedia [26]. Individual words were mapped to known embeddings and then fed into an embedding layer. We also used the Keras Bidirectional, GRU, Conv1D, GlobalAveragePooling1D and GlobalMaxPooling1D layers, effectively implementing a bidirectional recurrent neural network.

Results from classifying the *Topic* label based on abstracts with this alternative architecture can be seen in Table 12. When compared with experiment 2 results are either equal or slightly worse. It is reasonable to assume that the lack of training data is a contributing factor.

Discussion

A common issue with all experiments was lack of labeled data, which would definitely impact classification performance. Also, we only used the title or the abstract for features, which would impact some of the experiments. Beyond sparse data we know from conversations with NIPH that some of the *Topic* labels are impossible to infer from the publication title alone, which may explain the performance improvement

Table 12 Topic classification results from abstract (bidirectional RNN)

Class	Precision	Recall	F1
Diagnosis	0.78 (0.05)	0.58 (0.03)	0.67 (0.01)
Etiology	0.70 (0.03)	0.47 (0.06)	0.56 (0.03)
Experiences and perceptions; consequences; social, political, economic aspects	0.77 (0.11)	0.40 (0.05)	0.51 (0.02)
Infection prevention and control	0.53 (0.09)	0.19 (0.06)	0.28 (0.08)
Interventions targeted at system level	0.59 (0.07)	0.11 (0.08)	0.16 (0.12)
Interventions to treat the infected patient	0.76 (0.05)	0.29 (0.04)	0.41 (0.03)
Prevalence and incidence	0.54 (0.00)	0.28 (0.17)	0.34 (0.15)
Prognosis	0.67 (0.02)	0.55 (0.05)	0.61 (0.02)

in experiment 2 where the abstract was used instead of the title. At the same time, some of the other classes see no—or even a negative—boost to performance. The simple bag-of-words representation may be partly at fault. In addition, for some publications no abstract is available, even after augmenting with additional abstracts from the CORD-19 dataset. Finally, a key limitation is that we have no information on which parts of the abstract that lead the coders to make a coding decision. This makes the contribution of the abstract somewhat less precise. We do know, however, that the full text article has been consulted in cases where the coder was not able to make a decision from the abstract alone. We attempted to increase the vocabulary size to 10,000 words and observed some improvement to precision but typically at the expense of recall.

For experiment 3, classifying *Publication type*, we see that systematic reviews are much more likely to be both detected and classified correctly than non-systematic reviews; a reasonable explanation for this is that the former is much more likely to be explicitly named in the abstract than the latter, which is also a generic grouping category for “everything else”.

A similar observation can be made for experiment 4. As with the previous experiment, the *No data* class is likely to suffer from being implicit rather than explicit: from the reviewer’s point of view this label is applied in the absence rather than the presence of information.

A general source of error for all experiments is that the quality of the initial labeled publications is likely to fluctuate, especially for the earliest efforts. This is natural for any type of manual coding and classification project: it takes time for best practices to be established and knowledge to be disseminated among coders and the coding guide is likely to go through several revisions based on lessons learned during the coding process. Once more data is available this should become less of a problem.

When looking at ways to improve performance from a data point of view, an obvious activity would be to add additional training data. Since the evidence map project is still ongoing, additional coding data is being created and will provide a valuable basis for future experiments. Moreover, as the coders get more practice and experience the quality is likely to improve. Another possible effort is to improve the precision of the coding by having coders highlight the parts of the text that support their coding decision. This could improve the classifiers by allowing for more targeted

training. Since the full publication text has also been used, integrating full text where possible—or just indicating when this is the case—could make a positive difference. It is also worth noting that the code book specification has been simplified since our initial data export, which should make future classification easier.

Comparing the performance of our results with that of similar research on automated classification of scientific literature is not straightforward but some observations can be made. For example, in [21] we see F-scores of around 0.50 which is in the same area as our experiment 2, which had the largest number of classes. This study had a much larger training set but it is difficult to compare the complexity of the tasks. Often there are strict requirements that a high level of recall must be sustained, such as e.g. 0.95 in [18]. We have left considerations of what the acceptable recall—and the subsequent effect on workload reduction—for our task is for future work. In [19] the best F1 score was almost 0.90 but again with a more training data to work with.

Conclusions

We wanted to investigate if it was possible to use machine learning, specifically deep learning-based neural network models, to replicate a coding and classification scheme applied by expert coders over a period of several weeks to COVID-19-related publications. Our experiments showed that even with the simplest possible text representations and generic neural network architectures it was possible to get promising results.

To improve results further a natural place to start would be experimenting with deep learning architectures and best-practices that are better geared towards text classification, not the least when it comes to making use of word context and embeddings rather than the simple one-hot encoding currently employed. We conducted one experiment using external pretrained embedding vectors but they did not provide any performance boost. Further experiments with more data are highly relevant. Also, for small data sets traditional approaches such as support-vector machines often exhibit comparable performance to neural nets and thus warrants a comparison. Since both the evidence map project and the ensuing research collaboration came about in a rush we hope to address these improvements in future work. The aim of this paper is not methodological novelty but rather to highlight the potential of a unique handcrafted dataset.

The long-term goal is to build classifiers that can be used as a basis for coding process and decision support, thereby reducing the time spent and effort needed by the coders. While high classifier performance is a key requirement, the importance of user interface should not be forgotten. This is a particular challenge for applications where machine learning is a key component. The suggested codings will never be perfect and it is therefore crucial to establish a coder workbench that allows for both approving, modifying and rejecting the automated suggestions while at the same time allowing for manual review and oversight. Given that the coding will be an ongoing process, finding ways to iteratively improve the classifiers would be of particular interest. The history of health-related decision support applications is both long and chequered—but with several lessons to learn from [27].

For our experiments in this paper a simple evaluation against the gold standard was sufficient. However, when applying the classifiers towards assisting the coding process different evaluation metrics must be considered, e.g. the time spent coding and changes

to inter-coder agreement. Automated approaches are evaluated based on measurements of precision and recall but we know very little about what this means to those who wish to implement automation into workflows where a high-quality knowledge product is expected as the result. More knowledge is needed about acceptable error rates when applying decision-support technology to real-life needs and workflows.

While these initial experiments show promise for automated coding it is unlikely that the need for manual verification will be completely eliminated. Nonetheless, the amplification of highly skilled manual labor will improve quality, timeliness and frequency of updates by automating repetitive chores, new content detection, evidence integration, validation and consistency of results. Explainable artificial intelligence (AI) and explicit semantic reasoning in verifiable processes will allow experts to make predictable and trustable high-quality evidence maps. While we see immediate short-term potential in improving how knowledge is communicated for handling the COVID-19 crisis, the proposed technology can also have longer-term effects on medical information dissemination and management. As research communities become more advanced, global and specialised, the need for handling information flow and establishing best practices is unlikely to subside.

Abbreviations

AI: Artificial Intelligence; COVID-19: COVID-19 Open Research Dataset; JSON: JavaScript Object Notation; MeSH: Medical Subject Headings; NIPH: Norwegian Institute of Public Health; NLP: Natural language processing; NTNU: Norwegian University of Science and Technology.

Acknowledgements

The authors would like to thank NIPH for providing datasets and Thomas James at EPPI for useful background discussions.

Authors' contributions

TBR was the main author and performed the experiments. LS and ØN contributed on background and related work and the overall direction of the article, as well as facilitating the collaboration with NIPH. AEM and GEV provided data and contributed on background information. All authors have read and approved the final manuscript.

About this supplement

This article has been published as part of BMC Bioinformatics Volume 22 Supplement 11 2021: Proceedings of the 14th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO 2020). The full contents of the supplement are available at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-11>.

Funding

This work was partially supported by the BIGMED project funded by the Norwegian Research Council, Project No. 259055. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Publications costs are funded by NTNU.

Availability of data and materials

The data we used is not available yet as the project where it was collected is still ongoing. Release of a more comprehensive dataset will be considered for future work.

Declarations

Ethics approval and consent to participate

No ethics approval was required for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

²Reviews and Health Technology Assessments, Norwegian Institute of Public Health (NIPH), Oslo, Norway.

Received: 10 August 2021 Accepted: 23 August 2021
Published: 21 October 2021

References

- Glassiou PP, Sanders S, Hoffmann T. Waste in covid-19 research. *BMJ*. 2020. <https://doi.org/10.1136/bmj.m1847>.
- Norwegian Institute of Public Health. A systematic and living evidence map on COVID-19; 2020. <https://www.fhi.no/contentassets/e64790be5d3b4c4abe1f1be25fc862ce/covid-19-evidence-map-protocol-20200403.pdf>. Accessed 26 Mar 2021.
- OMara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev*. 2015;4(1):5.
- Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade AD, Wang K, Wilhelm C, Xie B, Raymond D, Weld DS, Etzioni O, Kohlmeier S. Cord-19: the covid-19 open research dataset. 2020. [arXiv:2004.10706](https://arxiv.org/abs/2004.10706).
- Oakley A, Gough D, Oliver S, Thomas J. The politics of evidence and methodology: lessons from the EPPI-Centre. *Evid Policy: J Res Debate Pract*. 2005;1(1):5–32.
- Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. 2018. [arXiv:1708.02709](https://arxiv.org/abs/1708.02709). Accessed 15 June 2020.
- Minaree S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning based text classification: a comprehensive review; 2020. [arXiv:2004.03705](https://arxiv.org/abs/2004.03705). Accessed 26 Mar 2021.
- Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*. 2019;29(2):86–101. <https://doi.org/10.1016/j.zemedi.2018.12.003>.
- Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. *J Med Internet Res*. 2018;20(6):10281. <https://doi.org/10.2196/10281>.
- Li M, Fei Z, Zeng M, Wu F, Li Y, Pan Y, Wang J. Automated ICD-9 coding via a deep learning approach. *IEEE/ACM Trans Comput Biol Bioinf*. 2019;16(4):1193–202.
- Obeid J, Heider P, Weeda E, Matuskowitz A, Carr C, Gagnon K, Crawford T, Meystre S. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Stud Health Technol Inform*. 2019;264:283–7. <https://doi.org/10.3233/SHTI190228>.
- Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B, Xu H. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc*. 2019. <https://doi.org/10.1093/jamia/ocz200>.
- Créquit P, Boutron I, Meerpohl J, Williams HC, Craig J, Ravaut P. Future of evidence ecosystem series: 2. Current opportunities and need for better tools and methods. *J Clin Epidemiol*. 2020. <https://doi.org/10.1016/j.jclinepi.2020.01.023>.
- Vandvik PO, Brandt L. Evidence ecosystems and learning health systems: why bother? *J Clin Epidemiol*. 2020. <https://doi.org/10.1016/j.jclinepi.2020.02.008>.
- Przybyla P, Brockmeier AJ, Kontonatsios G, Le Pogam M, McNaught J, von Elm E, Nolan K, Ananiadou S. Prioritising references for systematic reviews with RobotAnalyst: a user study. *Res Synthesis Methods*. 2018;9(3):470–88. <https://doi.org/10.1002/rsrm.1311>.
- Brockmeier AJ, Mu T, Ananiadou S, Goulermas JY. Self-tuned descriptive document clustering using a predictive network. *IEEE Trans Knowl Data Eng*. 2018;30(10):1929–42. <https://doi.org/10.1109/TKDE.2017.2781721>.
- Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. *J Biomed Inform*. 2016;62:59–65. <https://doi.org/10.1016/j.jbi.2016.06.001>.
- Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc*. 2006;13(2):206–19. <https://doi.org/10.1197/jamia.m1929>.
- Řehůřek R, Sojka P. Automated classification and categorization of mathematical knowledge. In: *Autexier S, Campbell J, Rubio J, Sorge V, Suzuki M, Wiedijk F, editors. Intelligent computer mathematics*. Berlin: Springer; 2008. p. 543–57.
- Joorabchi A, Mahdi AE. An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *J Inform Sci*. 2011;37(5):499–514. <https://doi.org/10.1177/016555111417785>.
- Romanov A, Lomotin K, Kozlova E. Application of natural language processing algorithms to the task of automatic classification of Russian scientific texts. *Data Sci J*. 2019. <https://doi.org/10.5334/dsj-2019-037>.
- Thomas J, Brunton J. EPPI-reviewer: software for research synthesis; 2007.
- Chollet F, et al. Keras; 2015. <https://keras.io>. Accessed 26 Mar 2021.
- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. *Tensorflow: A system for large-scale machine learning*. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016. pp 265–283. <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>.
- Chollet F. *Deep learning with python*. 2nd ed. Shelter Island: Manning Publications Company; 2020.
- Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Empirical methods in natural language processing (EMNLP)*; 2014. pp 1532–1543. <http://www.aclweb.org/anthology/D14-1162>.
- Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*. 2003;10(6):523–30. <https://doi.org/10.1197/jamia.m1370>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PAPER D

Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health



Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services

Thomas Brox Rost¹, Carolyn Clausen², Øystein Nytrø¹, Roman Kopusov^{3,4}, Bennett Leventhal⁵, Odd Sverre Westbye^{2,6}, Victoria Bakken², Linda Helen Knudsen Flygel⁷, Kaban Koochakpour¹ and Norbert Skokauskas^{2*}

¹ Department of Computer Science, The Norwegian University of Science and Technology, Trondheim, Norway, ² Regional Centre for Child and Youth Mental Health and Child Welfare, Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, ³ Regional Centre for Child and Youth Mental Health and Child Welfare (RKBU) Northern Norway, The Arctic University of Norway (UiT), Tromsø, Norway, ⁴ Sechenov First Moscow State Medical University, Moscow, Russia, ⁵ Department of Psychiatry, Division of Child and Adolescent Psychiatry, The University of California San Francisco, San Francisco, CA, United States, ⁶ Department of Child and Adolescent Psychiatry, St. Olav's University Hospital, Trondheim, Norway, ⁷ Department of Mental Health, Haukeland University Hospital, Bergen, Norway

OPEN ACCESS

Edited by:

Andrea Raballo,
University of Perugia, Italy

Reviewed by:

Laura B. Ramsey,
Cincinnati Children's Hospital Medical
Center, United States
Lorenzo Pelizza,
AUSL Parma, Italy

*Correspondence:

Norbert Skokauskas
norbert.skokauskas@ntnu.no

Specialty section:

This article was submitted to
Child and Adolescent Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: 20 May 2020

Accepted: 24 November 2020

Published: 15 December 2020

Citation:

Rost TB, Clausen C, Nytrø Ø,
Kopusov R, Leventhal B,
Westbye OS, Bakken V, Flygel LHK,
Koochakpour K and Skokauskas N
(2020) Local, Early, and Precise:
Designing a Clinical Decision Support
System for Child and Adolescent
Mental Health Services.
Front. Psychiatry 11:564205.
doi: 10.3389/fpsy.2020.564205

Mental health disorders often develop during childhood and adolescence, causing long term and debilitating impacts at individual and societal levels. Local, early, and precise assessment and evidence-based treatment are key to achieve positive mental health outcomes and to avoid long-term care. Technological advancements, such as computerized Clinical Decision Support Systems (CDSSs), can support practitioners in providing evidence-based care. While previous studies have found CDSS implementation helps to improve aspects of medical care, evidence is limited on its use for child and adolescent mental health care. This paper presents challenges and opportunities for adapting CDSS design and implementation to child and adolescent mental health services (CAMHS). To highlight the complexity of incorporating CDSSs within local CAMHS, we have structured the paper around four components to consider before designing and implementing the CDSS: supporting collaboration among multiple stakeholders involved in care; optimally using health data; accounting for comorbidities; and addressing the temporality of patient care. The proposed perspective is presented within the context of the child and adolescent mental health services in Norway and an ongoing Norwegian innovative research project, the Individualized Digital Decision Assist System (IDDEAS), for child and adolescent mental health disorders. Attention deficit hyperactivity disorder (ADHD) among children and adolescents serves as the case example. The integration of IDDEAS in Norway intends to yield significantly improved outcomes for children and adolescents with enduring mental health disorders, and ultimately serve as an educational opportunity for future international approaches to such CDSS design and implementation.

Keywords: child and adolescent mental health, clinical decision support system (CDSS), clinical decision support (CDS), innovation & technology strategy, child and adolescent psychiatry (CAP), child and adolescent mental health services (CAMHS)

INTRODUCTION

Nearly one half of mental health problems develop prior to the age 15 (1) and 75% of all psychiatric disorders have their onset prior to the age of 25 (2–4). In Norway, one out of five children has a mental disorder at any point in time (5, 6) and nearly five percent of all children and adolescents receive treatment in child and adolescent mental health services (CAMHS) (7, 8).

Modern electronic health records (EHRs) provide detailed documentation of a patient's health, but the complexity of psychiatric and neurodevelopmental disorders in childhood and adolescence requires clinical decision-making support beyond the EHRs' scope (9, 10). EHRs rarely provide adequate insight into the complex situations of psychiatric care, including recently updated biological frameworks for disorders and emerging methods for identifying syndromes (11–13). The incorporation of telepsychiatry and other computer supported health approaches can efficiently utilize existing resources to improve evidence-based early intervention and preventative CAMHS (13–15).

Clinical Decision Support Systems

A clinical decision support system (CDSS) aims to provide clinicians with real-time, step-by-step guidance through their clinical decision-making process (16–18). A CDSS intends to provide recommendations and guidance, not to replace the clinical judgment of practitioners. In general, a CDSS can be designed to rely solely on clinical practice guidelines to provide the evidence-based support, and/or incorporate previous patient cases by including healthcare datasets (18). The construction of guidelines for a CDSS is typically done with guideline development tools and computer-interpretable guideline (CIG) modeling languages, such as PROforma and SAGE (19). However, depending on the specific purpose of the CDSS, relying on modeled guidelines alone could be a suboptimal approach (20, 21). Traditional CDSS design and implementation aspects critical to successful CDSS adoption have included: (1) integration and adaptation to workflow; (2) construction of the information system structure and components; (3) knowledge management, interoperability, and sharing; (4) cognitive tasks and reasoning processes to be supported; (5) health system priorities and CDSS adoption paradigms; (6) quality improvement impacts, and (7) evaluation of effectiveness of decision support intervention (21, 22).

Child and Adolescent Mental Health Services in Norway

Norway is one of many Western nations that use an integrated approach for CAMHS. A family member, a teacher, or school counselor usually serves as the initial contact for children experiencing mental health problems, and refers them to a care provider. For example, if a teacher notices a child is challenged academically, they will involve the Educational and Psychological Counseling Service (PPT), which assesses the problem and determines whether special education assistance is an appropriate intervention, or if involvement of different

local, regional, or national services is most appropriate for the child (23).

Typically children are first referred to their local primary care provider (PCP) for further assessment. If the mental health problem is more complex in nature, a PCP needs to involve additional services from professionals who are trained to address such problems. For example, if there are child safety and well-being concerns, child protection services are involved, and if a child requires assessment and/or interventions by a child and adolescent psychiatrist, a referral to CAMHS is made (23, 24).

In addition to Norway's standardized, integral approach to patient assessment and treatment, the Norwegian Directorate of Health has also established national clinical guidelines and care pathways (i.e., Pakkeforløp in Norwegian) for several mental health disorders, similar to the United States' American Academy of Child and Adolescent Psychiatry (AACAP), formation of clinical updates and practice guidelines (25, 26). The national guidelines and standardized pathways help to improve the predictability and safety of care and facilitate collaboration between the different services involved (23, 27, 28).

CDSS DESIGN IN THE CAMHS CONTEXT

While CDSS implementation for general medicine has been well researched, the use of CDSS in CAMHS has been limited, with only a handful of studies focusing specifically on CAMHS, and many reporting shortcomings (11, 12, 18). CDSS design for CAMHS requires careful consideration of the complexity of the care process. The design and implementation should therefore take into consideration not only the previously documented challenges but also the structure and needs of local CAMHS (10–12).

To structure our discussion of the care context that a CAMHS CDSS must support, we have identified four key design considerations, representing (1) the collaborative aspect of mental health care, (2) the many and distributed sources of information, (3) the complexity introduced by multiple stakeholders and comorbidities, and (4) the long-term perspective of the care process.

A CDSS for Collaborative Care

Traditionally, standardized clinical guidelines and care pathways are designed for healthcare professionals directly involved in clinical care. But, providing quality care needs to involve all stakeholders, including teachers, community mentors (i.e., youth groups), coaches, as well as the patients and their families. Similar to clinical guidelines, traditionally CDSSs focus on the clinical provider and assists one individual through clinical decision-making (i.e., a psychologist or PCP) (22). There are several practical reasons for this, including legacy EHR systems' minimal interoperability, yet such approaches limit the scope of CDSS functionality, especially in CAMHS.

To maximize the value, usefulness, and impact of a CDSS, the correct information must reach all relevant stakeholders, whether directly or indirectly engaged (29). As the patient is the most important stakeholder in his or her own care, their active participation helps them to better understand the treatment,

and ultimately improves disease self-management (30, 31). In Norway, the Patients' Rights Act stipulates that all Norwegian citizens have the legal right to participate in their own care (32). Children and adolescents can provide consent and have a parent serve as a proxy (32, 33). Previous CDSS studies have shown CDSS system design should consider involvement of a parent as a proxy, as it increased patients' adherence to CDSS recommendations (17).

A CDSS for Application of Health Data

In a typical clinical scenario, decision-making is based on the patient's EHR, data from an associated patient database, and single-user data entry. The EHR should provide a holistic, comprehensive overview of the patient's health to maintain a consensus among all stakeholders involved in the patient's care. Assessment tools help identify the extent of a patient's problems and which stakeholders to involve in the patient's care. Self-reporting of symptoms has also become more common with the increased use and popularity of digital and web-based tools, especially among children and adolescents (34, 35). These methods of collecting information from multiple stakeholders involved, contributes to establishing a clearer picture of a patient's health. Clear communication and efficient sharing of the patient's health information is needed to provide the best quality care for each patient, as challenges with poor information flow and transparency directly affect the quality of care (36). A collaborative CDSS design, where multiple stakeholders participate in data collection and data entry, would increase the CDSS's utility as well as improve information flow among stakeholders (10).

Design of CDSS guidance based on analysis of health datasets has been found to provide greater improvement of clinical decision-making than guideline based CDSS suggestions alone (37). The data-driven approach to CDSS design can, not only provide decision-making support beyond the capacity of clinical guidelines, but also provide clinical learning opportunities (38). Reported secondary benefits of data-driven CDSS have included enhancing education, expanding research knowledge, improving guideline adherence, and clarifying training needs (39). Extending the role of a CDSS in this way can yield positive outcomes for patients with the most complex psychiatric needs.

A CDSS to Address Stakeholder Perspectives & Comorbidities

Applying a CDSS in clinical CAMHS also faces a challenge related to "cognitive collaboration" (40). "Cognitive collaboration" involves distributed cognitive processes from all stakeholders contributing to care, whose expertise covers a variety of professions (40, 41). Despite their common goal of helping the patient, the stakeholders' criteria for success, and their approaches to achieve that goal, may differ. For example, a school counselor's perspective on aspects of the clinical process may differ from that of a psychiatrist. A CDSS designed for one aspect of treatment might optimally address that particular focus, but this design approach could be less relevant to the overall clinical process if it neglects the "cognitive collaboration" involved in care (42).

In addition to multiple cognitive perspectives, the CDSS design also needs to account for comorbidities. Approximately 40% of all children and adolescents who meet the criteria for one disorder (i.e., anxiety, behavior, mood, or substance-use disorders) also meet the criteria for another disorder (43). Without considering abnormal symptomatic display or symptom overlap, comorbidity patterns can be concealed and mislead the practitioner to provide an invalid diagnosis (44). However, most CDSS models do not account for comorbidities, and research is scarce on how to apply multiple CIGs, in order to do so (11, 12, 45). A CDSS for CAMHS needs to be able to account for commonly occurring comorbidities, as well as the collaborative nature of clinical care (46).

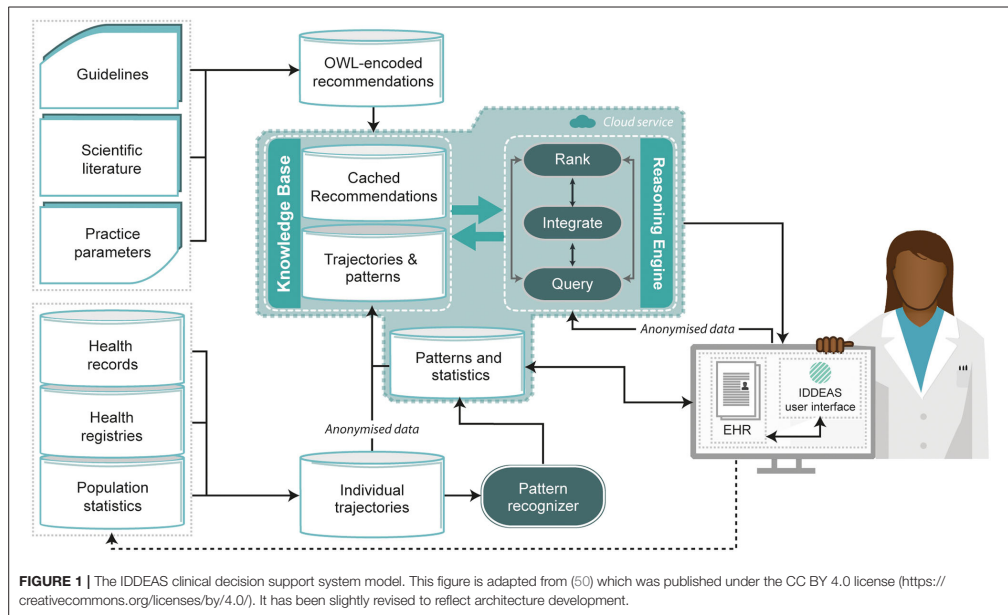
A CDSS for Temporality of Care

In Norway, the Patients' Rights Act guarantees every individual the right to immediate, appropriate care (32). For example, if a patient is referred to a psychologist or psychiatrist, they have the right to be seen within ten working days, and even sooner if the illness is deemed life-threatening (23, 32). Despite such policies, the patient's care progress and overall improvement of health can be delayed. Misdiagnosis, for example, can arise with child and adolescent mental health disorders due to the large variations in frequency, severity, and types of symptoms displayed, such as with ADHD (10).

Reaching a clinical diagnosis is only the first step in a complex and collaborative care process. A 2009 study on medical treatment for children with ADHD, found that only about half of the cohort managed to adhere to the ADHD medication plan (47). For a CDSS to be relevant to all components of the care process, potential complications that could arise in treatment management and follow-up also need to be taken into account. For example, a CDSS could be designed to consider any developments between appointments, or to register any irregularities prescribed medications and automatically alert the practitioner (48). To date, CDSS implementation and evaluations have predominantly focused on short-term outcomes rather than long-term care for the patient (42). While CDSS design has not yet optimally addressed the longitudinal and collaborative nature of patient care, many CIG modeling languages that can be used for a CDSS (i.e., EON, GASTON, etc.) do (49). In addition to utilizing a CIG language with longitudinal context, it is essential to assess how the different components of temporality of clinical care, and the specific timing of each intervention step, can impact the use of a CDSS (42).

A COMPLEX PROPOSITION TO MEET COMPLEX NEEDS: THE IDDEAS PROJECT

The complexities of a CDSS for CAMHS have all come under consideration in the development of the Individualized Digital DEcision Assist System (IDDEAS) project. IDDEAS, an innovation and research project in Norway, aims to design and implement a CDSS that can support the diagnoses and treatment of mental health disorders in children and adolescents, starting with ADHD as the first model clinical paradigm (50, 51). With



nearly 4% of all 12 year olds in Norway having ADHD at any point in time (7), the disorder will serve as the first case example for IDDEAS. IDDEAS brings innovation to patient care to allow earlier and more precise clinical decision-making.

The main goal of IDDEAS is to develop a CDSS that will improve mental health outcomes for children and adolescents by supporting the practitioner through clinical decision-making. IDDEAS specifically seeks to improve care by providing clinicians data-driven and evidence-based guidance in real time, to ensure earlier and more precise decision-making, avoid misdiagnosis and inefficient care practices, and improve individualized treatment management. In addition to the Norwegian CAMHS guidelines and clinical care pathways, IDDEAS will also use Norway's unique and existing resources—CAMHS datasets and other health datasets—to provide data-driven support.

The central and most important innovation in IDDEAS is the *Local Early and Precise (LEaP)* model, which allows for the application of IDDEAS *locally* in community settings, *early* in the clinical process, to add *precision* to patient care. The LEaP model is designed to provide real-time decision support for busy practitioners. IDDEAS integrates existing heterogeneous, geographically distinct, current and historical datasets, to generate new information and models to provide clinical decision support at the individual patient level (Figure 1). Data representing multiple episodes of care for different patients are structured into domains of inter-related concepts and

hierarchical clinical patterns. They are then ranked within the system, matched with the current patient and ultimately provided within the system's interface to support the practitioner through clinical decision making (50, 51). In addition, guidelines and other clinical recommendations are compiled and encoded before being combined with the data-driven trajectories and patterns to provide ranked suggestions in response to any practitioner queries (50, 51). By designing a CDSS that utilizes both guidelines and big data, the system has the potential to be curated based on evolving scientific evidence, and with the use of each individual patient's own EHR data to also build upon the available evidence base within the system (51).

The IDDEAS CDSS will be designed and evaluated in iterations. As this approach to CDSS design for CAMHS is relatively novel, to ensure IDDEAS is usable and appropriate for clinicians and patients, all iterations will be conducted collaboratively among the technical and clinical experts of the IDDEAS Consortium (50). With IDDEAS being an innovation project, each stage will build upon the previous one, with first identifying the needs of practitioners and assessing the perceived usability of the prototype system before going on to investigate the utility and efficacy of the system to care for real patients (50).

Preserving patient confidentiality is a fundamental project requirement. To mitigate the risk of re-identification we will seek to model patient trajectories in a way that reduces the patient representation to a set of care events (e.g., physiological

findings and health care system interactions). These will then be clustered so that we operate with representations of similar patient trajectories rather than unique trajectories tied to single individuals.

In developing the project, it was important first to consider the previously encountered challenges of successful CDSS implementations and then evaluate them within the context of the Norwegian local approach to CAMHS. A contribution of this paper is a framework to discuss which considerations a CDSS for local CAMHS must consider both for design and implementation: the involved stakeholders, how they share information, the explicit and implicit “cognitive collaboration” involved and how to address the longitudinal component of patient care. We recognize that some of these challenges, e.g., the handling of comorbidities or supporting multiple distributed stakeholders, are many-faceted and complex and do not often have straightforward solutions. In the IDDEAS project we seek to use this framework as a foundation for a structured engagement with our clinician partners and ultimately better understand the context and processes of CAMHS. We believe this will help us to understand the design and implementation trade-offs we must make but also where a CDSS can realistically have a positive impact on care delivery.

IDDEAS involves multiple stakeholders, including clinicians, researchers, computer engineers, service-user organization representatives, among others, and aims to facilitate “cognitive collaboration” throughout the project. While designated responsibilities lead to differing extents of active involvement from these stakeholders, the IDDEAS Consortium holds regular collaborative meetings for all stakeholders to consistently include multidisciplinary perspectives through development, evaluation and implementation. In addition to multidisciplinary cooperation, IDDEAS is nationally funded by the Norwegian Research Council (i.e., Norges Forskningsrådet) and involves collaboration on a national level (i.e., between different regional CAMH clinics), as well as on an international level, with Consortium members representing Norway, the United States, and several countries of the European Union (50, 51).

Overall, IDDEAS proposes an approach to CDSS design and implementation that not only utilizes the local available resources but also builds off of previously-established challenges and limitations of CDSS uptake and use in other settings, to try to avoid past shortcomings while adapting the approach to meet the local CAMHS.

REFERENCES

1. The World Health Organization. *Mental Health action plan 2013-2020*, WHO Library Cataloguing in Publication Data. Geneva: WHO, Department of Mental Health and Substance Abuse (2013). Available online at: https://www.who.int/mental_health/en/ (accessed May 20, 2020).
2. McGorry PD, Purcell R, Goldstone S, Amminger PG. Age of onset and timing of treatment for mental and substance use disorders: implications for preventive intervention strategies and models of care. *Curr Opin Psychiatry*. (2011) 24:301–6. doi: 10.1097/YCO.0b013e3283477a09

DISCUSSION

CDSS implementation in CAMHS has the potential to improve the quality of care and clinical outcomes for patients. The complexity of child and adolescent mental health requires a CDSS design that approaches treatment as a long-term, highly complex process. The optimal approach will encourage collaboration among stakeholders, involving their perspectives and knowledge as part of the foundation for the decision-making processes, while ensuring the patient receives appropriate, individualized care. The proposed IDDEAS in Norway offers helpful means to use innovative technology to improve CAMHS. While IDDEAS is first proposed for Norway, the project intends to test the CDSS within Scandinavia and Europe. A CDSS for child and adolescent mental health, designed and implemented based on established evidence, and using the LEAP approach, can result in improving the quality of services and the health of patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

TR was responsible for establishing the direction and writing the manuscript. CC also contributed substantially to the planning, writing, revising, and finalizing of the manuscript. TR, ØN, and KK contributed to the development of content on computer decision support systems and computer engineering. NS, BL, RK, LF, OW, and VB all contributed to the development of content related to clinical components. NS provided extensive feedback throughout the entirety of the manuscript's development process. All authors contributed to the article and approved the submitted version.

FUNDING

The IDDEAS project was funded by The Norwegian Research Council (grant no. 269117) and the Norwegian University of Science and Technology (NTNU). The Norwegian Research Council will provide funding in line with HELSEVEL (Programme on Health, Care and Welfare Services Research), which promotes integrated patient and user pathways and research and innovation activities aimed towards improving the quality of expertise and efficiency in health care services.

3. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB. Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry*. (2007) 20:359–64. doi: 10.1097/YCO.0b013e32816ebc8c
4. Kessler RC, Berglund P, Demler O, et al. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. (2005) 62:593–602. doi: 10.1001/archpsyc.62.6.593
5. NIPH. *Folkehelse rapporten 2014: Helsetilstanden i Norge*. Nasjonalt Folkehelseinstituttet (2014). Available online at: www.fhi.no (accessed May 20, 2020).

6. Det Kongelige Helse- og omsorgsdepartement. *Melding til Stortinget 19. 2014-2015 Folkehelsemeldingen: Mestring og muligheter*. Available online at: <https://www.regjeringen.no/no/dokumenter/meld.-st.-19-2014-2015/id2402807/> (accessed May 20, 2020).
7. NIPH. Quality of life and mental health among children and adolescents in Norway. In: *Public Health Report - Health Status in Norway*. Oslo: Institute of Public Health (NIPH) [updated (08.08.2019); read (28.10.2019)] (2019). Available online at: <https://www.fhi.no/en/op/hin/groups/mental-health-children-adolescents/> (accessed May 20, 2020).
8. HelseDirektoratet. *Aktivitetsdata for Psykisk Helsevern For Barn og Unge 2018*. Norsk Pasientregister, HelseDirektorat (2018). Available online at: <https://www.helseDirektoratet.no/rapporter/aktivitetsdata-for-psykisk-helsevern-for-barn-og-unge> (accessed May 20, 2020).
9. Leventhal B. What is the big deal about big data? *World Child Adolesc Psychiatry*. (2019) 17:3–6.
10. Cohen D. Assessing the effect of an electronic decision support system on children's mental health service outcomes. *J Technol Human Serv*. (2015) 33:225–40. doi: 10.1080/15228835.2015.1039687
11. Kopusov R, Fossum S, Frodd T, Nytro Ø, Leventhal B, Sourander A, et al. Clinical decision support systems in child and adolescent psychiatry: a systematic review. *Eur Child Adolesc Psychiatry*. (2017) 26:1309–17. doi: 10.1007/s00787-017-0992-0
12. Kopusov R, Frodd T, Nytro Ø, Leventhal B, Sourander A, Quaglini S, et al. Clinical decision support systems for child neuropsychiatric disorders: the time has come? *Ann Cogn Sci*. (2017) 1:12–5. doi: 10.36959/447/335
13. Skokauskas N, Fung D, Flaherty LT, Klitzing K, Püras D, Servili C, et al. Shaping the future of child and adolescent psychiatry. *Child Adolesc Psychiatry Ment Health*. (2019) 13:1–7. doi: 10.1186/s13034-019-0279-y
14. Zhou C, Crawford A, Sheral E, Kurdyak P, Sockalingam S. The impact of project ECHO on participant and patient outcomes: a systematic review. *Acad Med*. (2016) 91:1439–61. doi: 10.1097/ACM.0000000000001328
15. Wozney L, McGrath PJ, Gehring ND, Bennett K, Hugueta A, Hartling L, et al. eMental Healthcare technologies for anxiety and depression in childhood and adolescence: systematic review of studies reporting implementation outcomes. *JMIR Mental Health*. (2018) 5:1–20. doi: 10.2196/mental.9655
16. Osheroff JA, Teich JM, Levick D, Saldana L, Ferdinand T, Sittig DF, et al. *Improving Outcomes With Clinical Decision Support*. Chicago: HIMSS Publishing (2012).
17. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. *Implement Sci*. (2018) 13:114. doi: 10.1186/s13012-018-0790-1
18. Berner ES (ed), La Lande TJ. *Clinical Decision Support Systems: Theory and Practice*. Health Informatics. 3rd ed. Berner ES, editor. Cham: Springer International Publishing (2016). p. 1–17.
19. Khodambashi S, Nytro Ø. Reviewing clinical guideline development tools: Features and characteristics. *BMC Med Inform Decis Making*. (2017) 17:132. doi: 10.1186/s12911-017-0530-5
20. Ozaydin B. Data mining and clinical decision support systems. In: Berner ES, editor. *Clinical Decision Support Systems: Theory and Practice*. 3rd ed. Geneva: Springer International Publishing (2016). p. 45–68.
21. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. *J Biomed Inform*. (2018) 78:134–43. doi: 10.1016/j.jbi.2017.12.005
22. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. (2005) 330:765. doi: 10.1136/bmj.38398.500764.8F
23. HelseDirektoratet. *Mental health care in Norway: Psykisk helsehjelp/Mental health care (Ungdom)* (2008). Available online at: www.psykisk.no (accessed May 20, 2020).
24. HelseDirektoratet. *Adhd/hyperkinetisk Forstyrrelse – Nasjonal Faglig retningsslinje for Utredning, Behandling og Oppfølging*. (2019). Available online at: <https://www.helseDirektoratet.no/retningslinjer/adhd> (accessed May 20, 2020).
25. HelseDirektoratet. *Pakkeforløp: Psykiske lidelser - Barn og Unge*. (2019). Available online at: <https://www.helseDirektoratet.no/pakkeforlop/psykiske-lidelser-barn-og-unge> (accessed May 20, 2020).
26. American Academy of Child and Adolescent Psychiatry (AACAP). *Project ECHO in Telesychiatry with Youth* Dr. Ujjwal Ramtekkar. (2019). Available online at: https://www.aacap.org/AACAP/Clinical_Practice_Center/Business_of_Practice/Telesychiatry/Toolkit%20Videos/project_echo.aspx (accessed May 20, 2020).
27. Biringer E, Hartveit M. A future for pathways in mental health care in Norway: a discussion paper based in El-Ghorr et al. (2010). *Int J Care Pathw*. (2011) 15:18–20. doi: 10.1177/205343541110500104
28. Johansson KA, Nygaard E, Herlofsen B, Lindemarf F. Implementation of the 2013 amended patients' rights act in norway: clinical priority guidelines and access to specialised health care. *Health Policy*. (2017) 121:346–53. doi: 10.1016/j.healthpol.2017.02.007
29. Sirajuddin AM, Osheroff JA, Sittig DF, Chuo J, Velasco F, Collins DA. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support. *J Healthc Inf Manage*. (2009) 23:38–45. Available online at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3316472/>
30. Jimison HB, Gordon CM, Berner ES (eds.). *Clinical Decision Support Systems: Theory and Practice*. Health Informatics. 3rd ed. Cham: Springer International Publishing (2016). p. 163–79.
31. Koskela T, Sandström S, Mäkinen J, Liira H. User perspectives on an electronic decision-support tool performing comprehensive medication reviews- a focus group study with physicians and nurses. *BMC Med Inform Dec Making*. (2016) 16:1–9. doi: 10.1186/s12911-016-0245-z
32. LOVDATA. *Patient and User Rights Act. 2. juli 1999 nr. 63 om pasient- og brukerrettigheter, 1999*. Ministry of Health and Care Services (2020). Available online at: <https://lovdata.no/dokument/NL/lov/1999-07-02-63> (accessed May 20, 2020).
33. LOVDATA. *Act on Municipal Health and Care Services*. Ministry of Health and Care Services (2020). Available online at: <https://lovdata.no/dokument/NL/lov/2011-06-24-30> (accessed May 20, 2020).
34. Kobak K, Townsend L, Birmaher B, Milham M, Kaufman J. Computer-Assisted psychiatric diagnosis. *J Am Acad Child Adolesc Psychiatry*. (2020) 59:213. doi: 10.1016/j.jaac.2019.04.021
35. Grealish A, Hunter A, Glaze R, Potter L. Telemedicine in a child and adolescent mental health services: participants' acceptance and utilization. *J Telemed Telecare*. (2005) 11: S1:53–5. doi: 10.1258/1357633054461921
36. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*. (2003) 10:523–30. doi: 10.1197/jamia.M1370
37. Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices. *JAMA*. (2016) 315:562. doi: 10.1001/jama.2016.0275
38. Dagliati A, Tibollo V, Sacchi L, Malovini A, Limongelli I, Gabetta M, et al. Big data as a driver for clinical decision support systems: a learning health systems perspective. *Front Dig Human*. (2018) 5:1–7. doi: 10.3389/fdigh.2018.00008
39. González-Ferrer A, Valcárcel AM, Cuesta M, Cháfer J, Runkle I. Development of a computer- interpretable clinical guideline model for decision support in the differential diagnosis of hyponatremia. *Int J Med Inform*. (2017) 103:55–64. doi: 10.1016/j.ijmedinf.2017.04.014
40. Xiao Y. Artifacts and collaborative work in healthcare: methodological, theoretical, and technological implications of the tangible. *J Biomed Inform*. (2005) 38:26–33. doi: 10.1016/j.jbi.2004.11.004
41. Cohen T, Blatter B, Almeida C, Shortliffe E, Patel V. A cognitive blueprint of collaboration in context: distributed cognition in the psychiatric emergency department. *Artif Intell Med*. (2006) 37:73–83. doi: 10.1016/j.artmed.2006.03.009
42. Zikos D, DeLellis N. CDSS-RM: a clinical decision support system reference model. *BMC Med Res Methodol*. (2018) 18:137. doi: 10.1186/s12874-018-0587-6
43. Merikangas KR, He J, Burstein M, Swanson SA, Avenevoli S, Cui L, et al. Lifetime prevalence of mental disorders in US adolescents: results from the

D-8 ► PAPER D LOCAL, EARLY, AND PRECISE: DESIGNING A CLINICAL DECISION SUPPORT SYSTEM FOR CHILD AND ADOLESCENT MENTAL HEALTH

Rost et al.

Decision Support for Youth Mental Health Care

- national comorbidity study-adolescent supplement (NCS-A). *J Am Acad Child Adolesc Psychiatry*. (2010) 49:980–9. doi: 10.1016/j.jaac.2010.05.017
44. Caron C, Rutter M. Comorbidity in child psychopathology: concepts, issues and research strategies. *J Child Psychol Psychiatry*. (1991) 32:1063–80. doi: 10.1111/j.1469-7610.1991.tb00350.x
45. Peleg M. Computer-interpretable clinical guidelines: a methodological review. *J Biomed Inform*. (2013) 46:744–63. doi: 10.1016/j.jbi.2013.06.009
46. Tchong JE, Bakken S, Bates DW, Bonner III H, Gandhi TK, Josephs M (eds.), et al. *Optimizing Strategies for Clinical Decision Support: Summary of a Meeting Series. The Learning Health System Series*. Washington, DC: National Academy of Medicine (2017).
47. Pappadopoulos E, Jensen PS, Chait AR, Arnold EL, Swanson JM, Greenhill LL, et al. Medication adherence in the mta: saliva methylphenidate samples vs. parent report and mediating effect of concomitant behavioral treatment. *J Am Acad Child Adolesc Psychiatry*. (2009) 48:501–10. doi: 10.1097/CHI.0b013e31819e23ed
48. Kane-Gill S, Achanta A, Kellum JA, Handler SM. Clinical decision support for drug related events: moving towards better prevention. *World J Crit Care Med*. (2016) 5:204–11. doi: 10.5492/wjccm.v5.i4.204
49. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, et al. Comparing computer-interpretable guideline models: a case-study approach. *J Am Med Inform Assoc*. (2003) 10:52–68. doi: 10.1197/jamia.M1135
50. Clausen C, Leventhal BL, Nytrø Ø, Kopusov R, Westbye OS, Rost TB, et al. Testing an individualized digital decision assist system for the diagnosis and management of mental and behavior disorders among children and adolescents. *BMC Med Inform Dec Making*. (2020) 20:232. doi: 10.1186/s12911-020-01239-2
51. Clausen C, Leventhal BL, Nytrø Ø, Kopusov R, Westbye OS, Rost TB, et al. Clinical decision support systems: an innovative approach to enhancing child and adolescent mental health services. *J Am Acad Child Adolesc Psychiatry*. (2020). doi: 10.1016/j.jaac.2020.09.018. [Epub ahead of print].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Rost, Clausen, Nytrø, Kopusov, Leventhal, Westbye, Bakken, Flygel, Koochakpour and Skokauskas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

PAPER E

Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development



OPEN ACCESS

EDITED BY
Andrea Raballo,
University of Perugia, Italy

REVIEWED BY
Branko Aleksic,
Nagoya University, Japan
Flávio Dias Silva,
Federal University of Tocantins, Brazil

*CORRESPONDENCE
Norbert Skokauskas
✉ norbert.skokauskas@ntnu.no

SPECIALTY SECTION
This article was submitted to
Adolescent and Young Adult Psychiatry,
a section of the journal
Frontiers in Psychiatry

RECEIVED 31 August 2022
ACCEPTED 09 February 2023
PUBLISHED 23 February 2023

CITATION
Clausen C, Leventhal B, Nytrø Ø, Koposov R,
Røst TB, Westbye OS, Koochakpour K, Frodl T,
Stien L and Skokauskas N (2023) Usability
of the IDDEAS prototype in child
and adolescent mental health services:
A qualitative study for clinical decision support
system development.
Front. Psychiatry 14:1033724.
doi: 10.3389/fpsyt.2023.1033724

COPYRIGHT
© 2023 Clausen, Leventhal, Nytrø, Koposov,
Røst, Westbye, Koochakpour, Frodl, Stien and
Skokauskas. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Usability of the IDDEAS prototype in child and adolescent mental health services: A qualitative study for clinical decision support system development

Carolyn Clausen¹, Bennett Leventhal², Øystein Nytrø³, Roman Koposov⁴, Thomas Brox Røst³, Odd Sverre Westbye^{1,5}, Kaban Koochakpour³, Thomas Frodl⁶, Line Stien¹ and Norbert Skokauskas^{1*}

¹Department of Mental Health, Regional Centre for Child and Youth Mental Health and Child Welfare (RKBU Central Norway), Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway, ²Department of Psychiatry and Behavioral Neuroscience, The University of Chicago, Chicago, IL, United States, ³Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway, ⁴RKBU Northern Norway, UiT The Arctic University of Norway, Tromsø, Norway, ⁵Department of Child and Adolescent Psychiatry, St. Olav's University Hospital, Trondheim, Norway, ⁶Department of Psychiatry, Psychotherapy and Psychosomatics, University Hospital RWTH Aachen, Aachen, Germany

Introduction: Child and adolescent mental health services (CAMHS) clinical decision support system (CDSS) provides clinicians with real-time support as they assess and treat patients. CDSS can integrate diverse clinical data for identifying child and adolescent mental health needs earlier and more comprehensively. Individualized Digital Decision Assist System (IDDEAS) has the potential to improve quality of care with enhanced efficiency and effectiveness.

Methods: We examined IDDEAS usability and functionality in a prototype for attention deficit hyperactivity disorder (ADHD), using a user-centered design process and qualitative methods with child and adolescent psychiatrists and clinical psychologists. Participants were recruited from Norwegian CAMHS and were randomly assigned patient case vignettes for clinical evaluation, with and without IDDEAS. Semi-structured interviews were conducted as one part of testing the usability of the prototype following a five-question interview guide. All interviews were recorded, transcribed, and analyzed following qualitative content analysis.

Results: Participants were the first 20 individuals from the larger IDDEAS prototype usability study. Seven participants explicitly stated a need for integration with the patient electronic health record system. Three participants commended the step-by-step guidance as potentially helpful for novice clinicians. One participant did not like the aesthetics of the IDDEAS at this stage. All participants were pleased about the display of the patient information along with guidelines and suggested that wider guideline coverage will make IDDEAS much more useful. Overall, participants emphasized the importance of maintaining the clinician as the decision-maker in the clinical process, and the overall potential utility of IDDEAS within Norwegian CAMHS.

Conclusion: Child and adolescent mental health services psychiatrists and psychologists expressed strong support for the IDDEAS clinical decision support system if better integrated in daily workflow. Further usability assessments and identification of additional IDDEAS requirements are necessary. A fully functioning, integrated version of IDDEAS has the potential to be an important support for clinicians in the early identification of risks for youth mental disorders and contribute to improved assessment and treatment of children and adolescents.

KEYWORDS

clinical decision support system (CDSS), child and adolescent mental health services (CAMHS), children and adolescents, attention deficit and hyperactivity disorder (ADHD), usability

Introduction

Mental health is a key component of overall health. Mental disorders are amongst the most common and debilitating clinical challenges. For example, depression is one of the leading causes of disability worldwide (1). Furthermore, following the first year of the COVID-19 pandemic, the global prevalence of depression and anxiety increased by 25% (1). While all people are susceptible to developing mental health problems, children and teenagers are most vulnerable, with 75% of all life-time mental disorders having their onset in childhood and adolescence (2, 3). In addition, environmental factors are more likely to negatively impact the developing brain, increasing the risk for mental disorders in youth and children (1, 4). Despite this, access to and availability of timely CAMHS is limited (4). Without appropriate early interventions, children and adolescent mental health symptoms can evolve into potentially lifelong mental disorders, yet 70% of those experiencing mental health problems go without receiving appropriate care (1, 4–6). As part of routine health care, children and adolescents should, but rarely do, receive early assessments for risks associated with mental disorders (7, 8). Detecting and managing these risks as early as possible can help to reduce costs of services as well as societal costs, and ultimately, help alleviate the high demand for more complex treatment services (8, 9).

CAMHS expansion requires not only redistributed health budgets to allocate a greater share of funding toward mental health, but also investment in additional technological resources and mental health informatics (1, 4). Telepsychiatry or virtual reality (VR) exposure therapy exercises, for example, have proven to be effective mental health care (10, 11). Other health information technologies (HIT), such as clinical decision support systems (CDSSs), may have even more potential for service enhancement (12, 13). A CDSS is a tool designed to improve healthcare delivery by enhancing precision and timeliness of medical decisions through provision of support based on targeted clinical knowledge and patient health information (14). CDSSs are designed for various specific purposes, such as risk identification, diagnostics, and prescription management support (9, 12, 14). They can be developed to provide support with the use of clinical practice guidelines, as well as employing artificial intelligence (AI) to map

aggregated patient health record data, commonly referred to as “big data” (11, 15, 16). Big data analytics and mental health informatics using AI can provide evidence from multiple sources to allow for an aggregation of knowledge, account for multifaceted patient situations, and gain important insights for future approaches to care (16, 17).

Because of the challenge in juxtaposing normative clinical guidelines, with empirical evidence in the form of care patterns, developing a CDSS requires collaborative, multi-disciplinary efforts to ensure a cohesive balance between the technological innovation and the clinical workflow (12, 18, 19). Human computer interaction (HCI) and user-centered design (UCD) methods allow for simulated experimental and observational approaches that provide valuable insight into user workflow and clinician problem-solving needs. This process informs development, based on close collaborations with the end-users throughout innovation and research (13, 20–22).

Clinical decision support systems have found some significant success in general medicine and adult mental health but have yet to be adequately developed and implemented to CAMHS (18, 23, 24). The development of a CDSS for CAMHS faces systematic obstacles, including the lack of coordination amongst services and the limited accessibility of patient health data records used to develop a CDSS for CAMHS (25). While standardized clinical practice guidelines can be easily modeled for inclusion in a CDSS as part of an electronic health record (EHR) platform, providing decision support based on local practice patterns embedded in aggregated patient data can be challenging, as it requires access to hybrid and multi-source clinical data with approval from ethical committees and adherence to data protection regulations (i.e., General Data Protection Regulations-GDPR) alike (11, 14, 16, 20). Despite the challenges, the integration of health data has continued to exhibit potential for improving healthcare services (16).

Continued digital development, utilizing previously collected patient health data, has the potential to provide innovative solutions to acknowledge limitations within health services (4). With the digitalization of health services across specializations, integration of additional information and data from other information systems, could provide clinicians with transparent and holistic insight into a patient's current needs (14, 16). Exploiting all possibilities of digital solutions within a CDSS, not only limited to

patient health information from the EHR system but additionally encompassing digital case notes and hospital information systems, could provide a more efficient way to address the dynamics involved within CAMHS (11).

In Norway, the Individualized Digital Decision Assist System (IDDEAS) will be the first CDSS in CAMHS that uses both “big data” analytics and standardized clinical guidelines. Norwegian CAMHS are facing substantial increasing demand amidst the COVID-19 pandemic, like elsewhere in the world (26, 27). In 2021, almost 65,000 Norwegian children and adolescents received mental health care – a 14% increase from the previous year (26). Furthermore, over the course of the year nearly 36,000 referrals for mental health care have been reported for children and young people (26). The Norwegian National Association of Child and Adolescent Mental Health Services (N-BUP), established in 1958, has historically been responsible for providing a basis to connect all CAMHS in Norway and continuing to promote coordination and sharing knowledge amongst CAMHS (28). While N-BUP actively helps to facilitate the dissemination and sharing of important CAMHS information through research and management conferences annually, there is still invaluable CAMHS knowledge that has yet to be utilized- previously collected CAMHS individual patient EHR data (i.e., BUP-data) (29). The previously established EHR system of BUP-data was the first of its kind in Norway to be able to provide data comparisons on an individual patient basis (29). While the EHR system has been replaced, utilizing the knowledge acquired within BUP-data, in combination with standardized clinical practice guidelines, has the potential to provide Norwegian CAMHS with additional support to meet the mental health needs of children and adolescents (30). Upon receiving access to this invaluable resource, with support from N-BUP, and in close collaboration with its’ clinicians, the IDDEAS project is developing and researching a CDSS to provide clinicians in Norwegian CAMHS with real-time decision support, in part by BUP-data, but also with standardized clinical guidelines, including DSM-5 and ICD-10 (11).

The IDDEAS prototype is in the process of formative usability testing, including this qualitative study. This study aimed to understand CAMHS clinicians’ overall perceptions of IDDEAS prototype usability while also examining potential barriers to implementation and specific needs to be met in the development of the CDSS. The objectives of this study were to 1) explore clinicians decision-making processes; 2) investigate the perceived usability and functionality of the IDDEAS prototype; and 3) identify the user-perspectives on IDDEAS, to inform continued development and feasibility within Norwegian CAMHS.

Materials and methods

Study design

This is a mixed-methods study to evaluate IDDEAS, a decision support system for diagnosis and treatment of children and adolescents in Norwegian CAMHS. The IDDEAS project is organized into the following stages: (1) The Assessment of Needs and Preparation of IDDEAS; (2) The Development of the IDDEAS CDSS model; (3) The Evaluation of the IDDEAS CDSS; and (4) Implementation and Dissemination (see Figure 1). This qualitative

study reports on the interviews conducted as one component of the usability evaluation of the first IDDEAS prototype (11).

This evaluation process utilizes user-centered design (UCD) methods, with the testing of the CDSS conducted in phases of developmental iterations. The UCD methods include formative usability sessions (12, 31), cognitive walk-through/think-aloud procedures (5, 32), iterative development with end-users, and utilization of both qualitative and quantitative methods of inquiry (31, 33). As part of UCD, the iterative development of the CDSS involves continuous collaboration with CAMHS clinicians. The specific methods and the development plan are detailed in the IDDEAS project protocol (11). The present study serves as the first usability test, using UCD methods to investigate Norwegian CAMHS clinicians’ perceptions of the usability, utility, and overall functionality of the IDDEAS prototype.

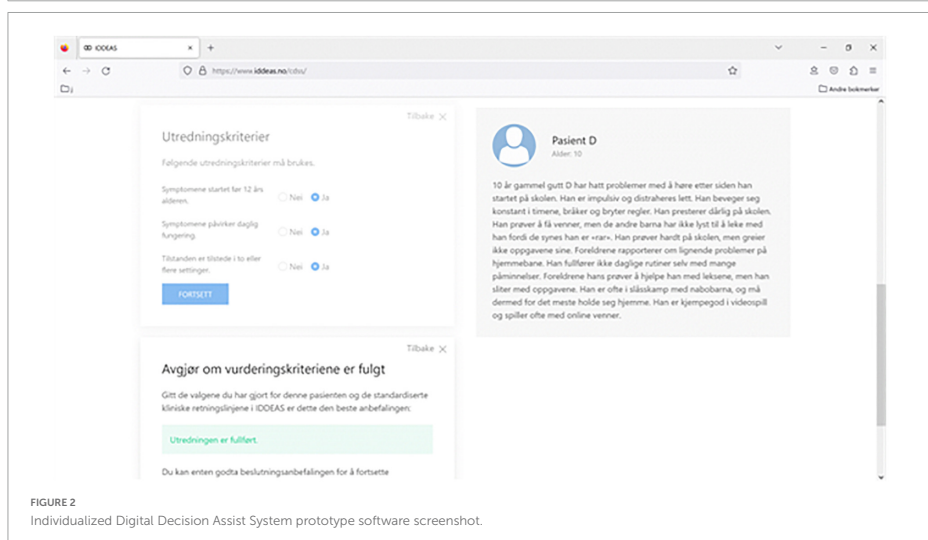
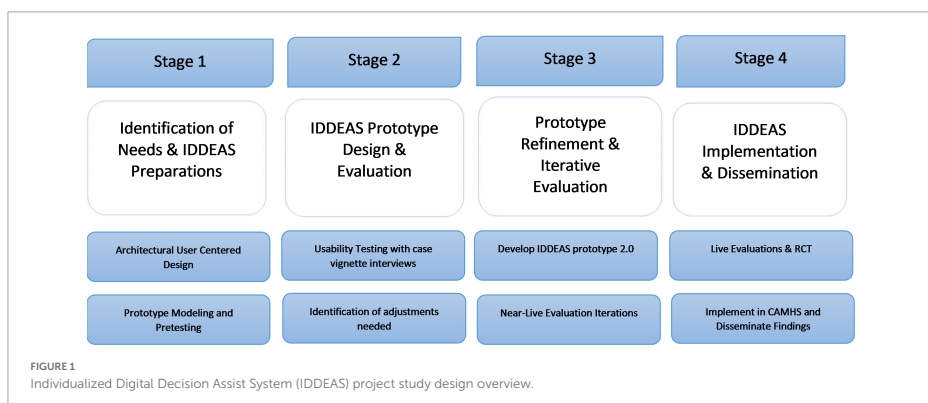
IDDEAS prototype

The IDDEAS prototype allows for exploration of the ability of IDDEAS guidelines to provide decision support for diagnosis and treatment of attention deficit and hyperactivity disorder (ADHD) (see Figure 2). ADHD is a neurodevelopmental disorder characterized by inattention, hyperactivity, and impulsivity, ultimately causing impaired functioning for the individual (9). The IDDEAS prototype at this stage uses ADHD as the first clinical model paradigm. Preparation of IDDEAS includes the validation of the clinical materials and the user-interface. The IDDEAS guidelines were previously validated by the IDDEAS clinical research team using the DSM-5 and ICD-10 criteria. Focus groups were used to pre-test content prior to the IDDEAS prototype evaluation.

Each IDDEAS prototype evaluation session included having a clinician participant complete a concurrent, cognitive walk through/think-aloud procedure, as they critically appraised hypothetical patient case scenarios developed from real cases within CAMHS. A total of 20 patient case scenarios were collaboratively designed and validated by the IDDEAS team (BL, NS, RK). Out of the 20 possible cases, each participant was randomly assigned four to assess, two of which were to be assessed while using the IDDEAS prototype (ADHD modeled guidelines) and two without. Use of IDDEAS was similarly randomly assigned. Throughout the assessment of the four cases, participants were asked to follow a think-aloud procedure and provide a concurrent walk through of the clinical procedure they would follow if the patients were real. They were also asked to provide additional patient information they perceived to be potentially necessary to complete their clinical assessment. Finally, participants were asked to provide their overall perceptions of the IDDEAS prototype and its usability, functionality, and potential utility.

Setting and sampling

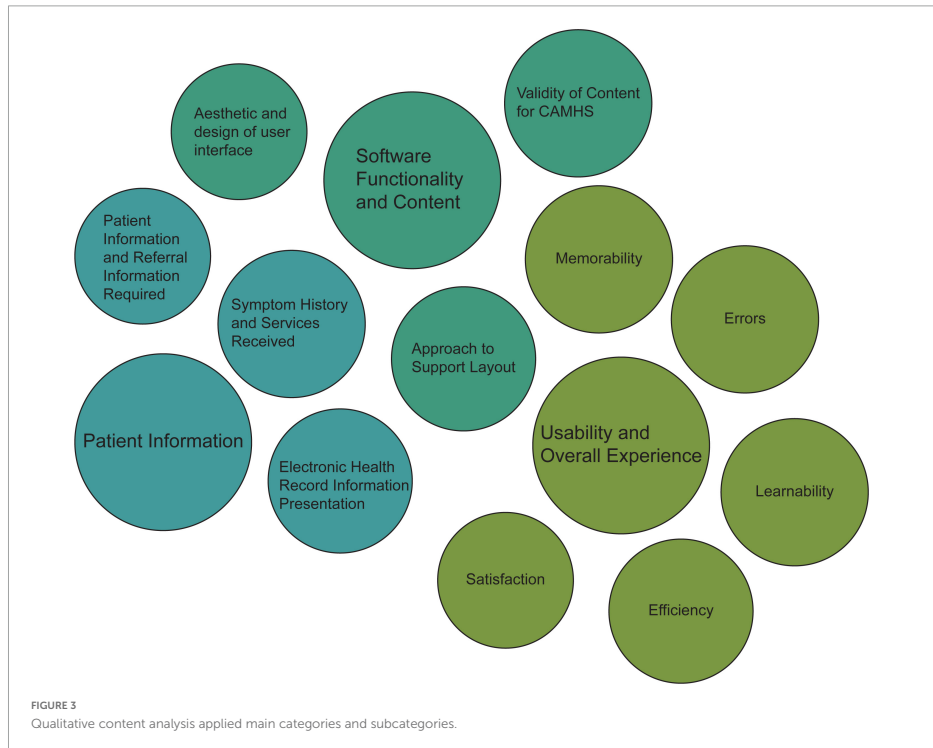
The participants ($n = 20$) were those who first participated from the larger cohort evaluation of the IDDEAS prototype (11). We (CC) directly contacted all potential participants who had been recommended by N-BUP board members and those from



a random list of service providers. To promote privacy and confidentiality, an invitation email with background information about the IDDEAS team and consortium, as well as the project's scope and aims, was sent to each potential participant. We (CC) met with each participant prior to the evaluation session in order to go through the proposed study procedure, as well as to provide participants with an opportunity to get acquainted and ask any potential remaining questions. Upon agreeing to participate in the study, each participant created their own profile on the IDDEAS portal and in accordance with the Norwegian Centre for Research Data (NSD) protocol, completed the informed consent process. Initial focus group discussions and pre-testing sessions were conducted beginning in March 2020, with the interviews taking place until Spring 2022.

Research instrument

A semi-structured interview guide with five questions was developed collaboratively by the IDDEAS team, based on the specific research question and the overall objectives of the IDDEAS project. The interview guide was created following the Mayring qualitative content analysis (QCA) approach (34) and is similar to those implemented by Schaaf et al. (12) and Baysari et al. (30). The final interview guide was confirmed by the IDDEAS team and translated, making it available in both English and Norwegian (see [Supplementary Appendix 1](#)). We (CC) conducted preliminary internal testing with members of the IDDEAS team. After the internal testing, a small focus group interview was held with four Norwegian CAMHS psychologists and psychiatrists who all



met the inclusion criteria for the qualitative study. Participants were deemed eligible for inclusion if they were either a child and adolescent psychiatrist or psychologist. All potential participants who did not meet the inclusion criteria were excluded. Study participants were given the option to choose to complete their interview in English or Norwegian.

Data collection

The interviews took place at the end of the IDDEAS prototype usability evaluation sessions. The study was conducted following UCD methodology and standardized criteria for qualitative research, including the consolidated criteria for reporting qualitative research (COREQ) and the Standards for Reporting Qualitative Research (SRQR) (35). The first author (CC) was responsible for interviewing the participants. The research question, interview guide, and qualitative data categorization system were all developed by CC and verified by the IDDEAS team.

After completion of informed consent and establishing a profile with the IDDEAS portal,¹ participants were invited to

meet with CC, either in person or online. Due to COVID-19 meeting regulations and safety requirements, all invitations sent out were *via* the Microsoft Teams online platform. All interviews were recorded and transcribed, word-for-word. All interviews were conducted directly following the completion of the IDDEAS prototype assessment's case appraisal procedure. All interviews took place within one session and no interviews were repeated or redone. All transcripts were saved within a secure, password protected zip file and stored on the Norwegian University of Science and Technology (NTNU) secure server in preparation for data analysis. No personal or sensitive data was included in accordance with the Norwegian Centre for Research Data (NSD) protocol requirements and research data management permission granted (reference code: 100166).

Data analysis

In line with QCA methods, a category system and coding rules were developed for the qualitative data analysis. The system was based on the research question and the study's objectives, with the specific categories developed to determine which textual passages to take into consideration. Following an

1 www.IDDEAS.no/

inductive category development procedure, the categories are tentative and deduced step-by-step, as applicable. The proposed categories were presented to the IDDEAS team members for theoretical structure verification prior to application to data material and formative/summative checks of reliability. Theoretical based definitions, examples of applicable text passages, and coding rules for each category, were collated within a coding agenda (36). As suggested by Mayring (34, 37), falling within the range of 10–50%, 35% of the transcribed material was checked with the preliminary categories and assessed for adequate representation of theoretical foundation and encompassing the text content. The proposed categories and the coding agenda were presented to the IDDEAS team and underwent revision before completing data analysis. The categories were revised from three main categories and 12 subcategories in Version 1 to a total of 11 subcategories in Version 2 (see [Supplementary Appendix 2](#) for more details). The final category system consisted of three main categories and eleven subcategories (see [Figure 3](#)).

All text passages from the interview transcripts were extracted and organized following the deductive category application model (34). The content-analytical coding rules were followed, to keep the process of category application as controlled as possible and to determine the most appropriate category. If there was a text passage that could not be assigned to a category, this was discussed with the IDDEAS team. After assigning all text passages to categories, all included within each category were then summarized and an example quotation was extracted for representation of the content. The extracted quotations that best represented the content of the category were chosen to represent the main findings. Any quotations in Norwegian were translated to English.

Results

Participants

The participants represent ten CAMHS clinics. Most participants identified as men ($n = 11$) with the rest identifying as women ($n = 9$). Fourteen were CAMHS psychologists, while 6 were CAMHS psychiatrists. The participants had varied experience working in CAMHS: no participants worked in CAMHS for less than 6 months, one participant had worked in CAMHS for 6–12 months; two participants worked in CAMHS for 1–4 years, and 17 reported to have worked in CAMHS for 4 years or more. The IDDEAS prototype evaluation session duration ranged from 30'20" to 93'51" with the 5-question interview mean duration of 6'45" ranging from 1'15" to 11'34".

Main results by category

The following sections present the results organized by the deductive QCA categories. We provide example comments from each category. Three main categories were extracted: (1) Patient information, (2) Software Functionality, (3) Usability and Overall Experience (see [Figure 3](#)).

Category 1: Patient information

Patient information and referral information required

Most of the participants reported concern about insufficient patient information available during the evaluation session. Participants who raised this issue acknowledged that they understood that the evaluation procedure was to intentionally include hypothetical patient case scenarios with limited clinical data, as well as the limited ability to engage with the IDDEAS prototype at this stage in its development. One participant noted:

"[...] well, I thought with very limited patient information it makes it a little more difficult. But in real life setting I think it is a valuable clinical tool."

Participants shared that insufficient patient information made it difficult to arrive at one diagnosis and indicated a need for more information in order to adequately utilize suggestions. Participants acknowledged they were missing important information, such as what is currently established about the patient, and the ability for the information to be adjusted accordingly to keep up to date. One participant said:

"I think it could be clearer what is missing and why does the patient not fulfill ADHD criteria so that I can think critically about it and whether there is something that I missed. To have that structured at the end when I'm finished [...] because these symptoms are missing, for instance. So, if I am a bit unsure I can think about it."

Finally, it was not easy for participants to speculate how it would be to use the system in the future to input their own patient information changes or adapt to changes in patient data between sessions. It was stated that decision support could be very useful when additional referral information is available and, for junior colleagues, guidance on how to find missing information could be beneficial. One participant offered the following suggestion:

"I guess a good thing would be if it was a patient I knew but then a question I didn't know and it'd say I have to get this later and then the score would tell me something based on that- where you have the ability to get a score with the "provisional score" or notifying you that you haven't answered all of them yet."

Symptom history and services received

All participants found it important to have information about the patient's previous symptoms, the diagnostic/treatment history, and any previous services. Participants commonly explained that there was often insufficient information available on the family context and the relationship with parents. For the clinicians to feel they can adequately assess the current patient information, there is a need for thorough presentation of patient history and any collaborative services accessed that inform the patient's assessment. Multiple participants explained the importance of always having multiple hypotheses for patients, without knowing all services received already. More specifically, participants noted that while

there may be patient symptom history available, information from additional services involved in the care was not adequately presented. For example:

"[...] while there was a lot of information about the single individual, there was less about the context of the family [...] so I feel like having more of the child's contacts, as student, etc."

Electronic health record information presentation

All participants reacted positively to the presentation of the patient information directly adjacent to the guideline support. Participants noted the importance of being able to navigate between the guidelines and the patient information seamlessly, and being able to track location in the guideline, while maintaining access to patient information visible on the other side of the screen. One participant explained:

"If it was something I thought I knew but when I read it again, I don't actually know it. So, it is very helpful to have those things next to each other and be reminded of specific criteria so you can systematically see where you are at (for the patient). So, I liked that."

Participants noted that not only is displaying the patient information and criteria side-by-side advantageous, but it could be potentially important to have the patient EHR data integrated with the decision support in the future. For example:

"[...] if I have a patient, should I then write in all of the symptoms or the case? The intention will be that you will have the EHR and they will be within (integrated)? You might be receiving alerts, etc. I think it would be very good to receive reminders."

Category 2: Software functionality and content

Validity of content for CAMHS

Participants' perceptions of the IDDEAS prototype functionality varied. Some participants had problems interpreting guideline content; they did not like the phrasing and found instructions difficult to understand. Some participants questioned future functionality of IDDEAS with the prototype guidelines requiring the participants to click through all guidelines support materials, regardless of whether they might need to review that specific information or not. For example, one participant stated:

"[...] I think point 2 is obvious, point three as well actually [...] and point 4 as well, I guess. The only point that might be helpful is point 1."

While this participant was one who seemed hesitant about the need for providing guideline support throughout the clinical process, others spoke highly of the fact that the IDDEAS ADHD guidelines were detailed and encompassed all components of

standardized guidelines. Overall, most participants were pleased to be provided with guideline support that matched what they would instinctively do in their current practice and took comfort in knowing it would be available as novice clinicians might need explicit step-by-step guideline support through their assessment. As one clinician explained:

"[...] I see the diagnostic criteria and I am quite fast/or it quickly is matching with my expectations for what this is. So, I would be more concerned or more skeptical if it was a mismatch with my clinical experience or my knowing of what ADHD criteria are. So, it is logic in that sense."

With another participant explaining the potential benefits for novice clinicians by stating:

"I think IDDEAS is a tool that could be very useful, I think. Especially for young clinicians, [who are] not very experienced and having a system that you click and go further and see those symptoms, the criteria are there and then it helps in a decision making."

Aesthetic and design of user interface

In terms of the aesthetics and design of the IDDEAS prototype, all but one participant found the prototype to be adequate. These participants reported that the text was easy to read, and the guidelines were easy to use. The participant who did not find the aesthetics and design to be adequate noted that the IDDEAS prototype was not aesthetically pleasing due to being too much like a webpage. Most participants emphasized the simplicity as a positive design element. One participant stated:

"It was neutral. It just felt neutral. And that is alright because I don't think it needs to be a visually stimulating experience. But it's good that there are not many distractions, it is good that it quite clear and clean in a way."

Approach to support layout

Several clinicians reported trouble with the decision tree guideline format which requires "Yes" or "No" responses when criteria are met, or not. "I do not know" option was also voiced, as illustrated by a statement below.

"I missed the "I don't know" button, but except for that it was really clean."

Another participant explained further:

"[...] when I work in a field where the children's situations are so complicated, I don't really want to be guided to a yes or no this early in the process. So, there was something there that I didn't like so much."

However, most felt positive about the decision tree as it helped them structure their thought process and identify components of

Clausen et al.

10.3389/fpsy.2023.1033724

the guideline support that they liked and other aspects that could be changed for the next IDDEAS version. Participants found that the guideline decision support layout, in step-by-step, informative guideline criteria boxes, helped to structure the clinical assessment process:

"I liked how, or realized that the further I went, it helped me to organize in a sense, instead of just blurting out everything I thought, it was more structural in a sense [...]"

While this participant found the guideline-based support helped structure their efforts, they also noted that this layout could also negatively impact their work, if they were unable to track their progress in applying the guidelines; they feared losing track of progress if they closed one guideline box.

Participants also mentioned a significant concern about their inability to move directly to guideline criteria to allow for investigating differential diagnoses (i.e., investigate inattention criteria met instead of assessing for hyperactivity) rather than going through the entire ADHD guideline following along with the predetermined sequence of the guideline decision support boxes provided, one-at-a-time. Participants were also clear about the need to expand clinical guidelines beyond ADHD in order to address comorbidities and appropriately address symptoms commonly displayed across multiple disorders. This issue also suggested the need to have multiple guidelines and criteria available to allow for navigation from general to specific components of diagnostic criteria. One clinician explained:

"One thing I would like is access to all of the guidelines, whenever I want. Because what is going on in my mind is several hypotheses at the same time, and that is what I am appointed and educated to do, a differential diagnostic assessment..."

Category 3: Usability and overall experience

Satisfaction

Most participants indicated that they were not entirely clear about the potential usefulness and helpfulness of limited (to ADHD only) IDDEAS. However, most were hopeful and intrigued by IDDEAS and were interested in its potential even though the prototype had limited utility, as they could not use it in an interaction with "real patient" information.

Despite limitations of the current prototype, it was judged to be easy to use and participants were interested in seeing the ongoing developments. It seemed clear to all that IDDEAS' usefulness will increase with the expansion of the diagnostic decision tree and the ability to see whether patient symptoms lie along a threshold. A participant stated that they liked the tool because it helped them to structure their thoughts about the diagnosis. One participant reflected on the ease of use and user-friendliness specifically:

"It was very user friendly, actually. It was very intuitive and very easy [...] you know normally I wouldn't really think too much

about such things and that's probably a good thing, which means then it was probably fairly easy to move around inside. . .] I think it was decent."

Learnability

Learnability in this context refers to the ability to learn how to use the IDDEAS prototype. There were mixed thoughts about the "learnability" of IDDEAS. Participants noted that IDDEAS is intuitive, and there is potential for improved ease of use and helpfulness based on the positive degree of learnability. While some reported initial challenges, it did not take too much time to understand how to go through the system. One clinician explained that they would enjoy learning how to interact with the system in the future, over the current approach to clinical care:

"I think it was very useful. Like I can see myself finding it more fun to do these evaluations, like it reminded me kind of some sort of game or it's more pleasing to just look up in the EHR platform and papers and ICD manuals and stuff, if you know what I mean."

There were some barriers to the learnability of IDDEAS due to the user interface. More specifically, some participants specified that they found it difficult to use and interpret the prompts. One participant explained:

"[...] it was kind of easy to follow where you should be looking, with the exception of the red and green (buttons) [...]"

Another elaborated further, to explain:

"[...] Sometimes it can be like, okay there is a window there and there, and where do I start or what is most important to read first? [...] But I know that a lot of people that I work with are maybe kind of "tech hard" so having a very simple button with "start with this" because we have so many things to think about all the time and other distractions."

Some participants explained that when there was too much going on within the layout of the interface, it can be challenging. Participants suggested that adjusting the symbols indicating where to click and the wording used in the notifications could make it easier to learn. One stated:

"There was something that I had to click back and the X symbol, so (indicating) now quitting everything and then nothing back saying "leaving" or things are saved. The wording or the icons need to make it clear that okay I've completed this now [...]"

Efficiency

The efficiency of using the IDDEAS prototype was discussed both in terms of the current approaches to guideline provision and the potential improvement with developments. Participants found it hard to assess the efficiency of IDDEAS at this time, largely due to the limited capacity of the prototype. Participants discussed that

without seeing the entire program, it is difficult to fully appreciate the actual potential for IDDEAS and its contribution to practice efficiency and quality improvement.

One participant noted that they found it difficult to determine IDDEAS to be usable and useful at this stage due to the phrasing of the alerts in guideline boxes causing some delay. For example, understanding the intention of the “decline/accept recommendation” support message provided within the guideline and becoming acquainted with what exactly this was prompting them to do throughout their patient assessment procedure. Additionally noted was the requirement to click through each guideline support box and all criteria included within the ADHD guideline, negatively impacting the efficiency of their clinical procedure. One participant explained:

“So my mindset is more on speed and efficiency, and this is slow. It is slowing me down [...] and this is more like reading a book, so it is maybe actually more efficient to use the real book.”

Memorability

Memorability in this study refers to the ability of the user to remember the task at hand and the components involved in the procedure. Overall, participants spoke positively of the ability to follow the workflow to assess a patient while using the IDDEAS, even though this may be perceived differently from clinician to clinician, particularly based on their experience. One participant explained:

“I see especially with new psychologists that I have to make them okay with not knowing all the time and to be curious or uncertain and IDDEAS can help with this by widening the focus at the beginning and then narrowing it down as you go.”

Errors

Participants reported at times having encountered errors with the guideline support (i.e., “Not Supported” message displayed upon acceptance of a recommendation) and the navigation buttons (i.e., inability to close one guideline box without exiting entirety). Additionally, participants specifically discussed encountering glitches with the system generating repeat guideline boxes, the inability to access specific criteria when clicking yes in response to prompt suggestion, or falsely notifying the participant that the guideline is over when they have selected to reject the recommendation and continue their assessment. Participants specified that the errors encountered with the prototype made them find it less usable and appropriate at this stage of development. One participant explained in reference to the inability to access more of the guideline upon declining the guideline recommendation:

“Going into the project, I am probably on the side of being a little bit skeptical already based on the diagnostic system is trending toward categorical systems regarding children’s health and functioning, so I am probably a little bit difficult to convince regarding the usefulness of such a system...”

On the other hand, another clinician stated simply in reference to a glitch encountered:

“Okay besides the glitch [repeated guideline support box] if this is refined it could be very interesting tool, absolutely.”

Discussion

This study represents the first phase in the development of the IDDEAS CDSS. It is a qualitative study of how CAMHS clinicians perceived the usability of the IDDEAS prototype. As IDDEAS is developed iteratively and in collaboration with the end-users, revisions and adaptations are expected. This qualitative study provides valuable initial information about the usability of IDDEAS, while also identifying needs based on input from potential end-users, CAMHS clinicians.

Our study suggests that the first IDDEAS CDSS prototype needs to be adapted and adjusted to be perceived as usable and helpful. However, more importantly, there is a consensus amongst stakeholders that there is great potential for its usefulness with further development, as well as an eagerness for engagement in helping to inform the future development of the IDDEAS CDSS.

Clinicians were able to use the simulated explorative procedure to evaluate the usability of the IDDEAS prototype, and the potential for useful and helpful future versions of IDDEAS. Our experimental procedures allowed the clinicians to reflect on IDDEAS and suggest what could be better or different. While there was limited patient information and an inability to interact with a fully formed CDSS, the study allowed for us to learn about clinicians’ preferences with respect to what they need and do not need from the CDSS in CAMHS and EHR integration (38). Similarly, other CDSS development studies that have used hypothetical case scenarios found similar limitations dependent upon the state of the CDSS prototype but still identified important takeaways for the systems further development, including close integration of the patient information from the EHR (31). The main consensus elicited from our findings was the importance of quickly being able to identify patient information that is missing at the time of assessment. In this case, clinicians specified that with growing demand for services it is important to be able to efficiently determine whether a patient referral to CAMHS might be rejected or accepted. Furthermore, with global pandemics seemingly becoming a global societal norm, improved timeliness, and overall efficiency of the provision of care within CAMHS could potentially greatly benefit from further incorporation of well tested and validated HIT, such as a CDSS, as long as it is developed in accordance with end user needs.

Based on our findings, a guideline-based decision support system was helpful, but it needs to be able to provide clinicians with customized suggestions as to which clinical guidelines to reference, based on changes in the patient’s health status as well as services previously accessed. Interacting with the guideline-based support provided clinicians an opportunity to reflect on what they feel is lacking in the platforms currently used in CAMHS and speculate how IDDEAS could help to meet these needs in the future. It is also important to acknowledge that the guideline functionality serves as only one component of the overall functionality of

TABLE 1 Individualized Digital Decision Assist System (IDDEAS) prototype attributes: Perceived strengths and limitations.

Attribute	Perceived strength	Perceived limitation	Proposed development
The “accept recommendation” guideline support box	Shows the clinician that they are still in charge	Can be unclear for some clinicians how recommendation comes about	Show summary confirmation of what the clinician selected and optional box for where recommendation came from
“Criteria not fulfilled” notification guideline support box	Allows for the clinician to see what they know and what they do not know	Can be unclear based on phrasing of information in box	Provide notification for recommendation with evidence optional to access (BUP-data statistics-regional/department examples); simplify wording and appearance to be clear
Structured layout of guideline support boxes	Provides clinician with reminders for the important information to acquire within the assessment steps and following the structured diagnostic process ensuring reliable and standardized diagnostic procedure	Inability to navigate through the guideline outside of the step-by-step structured support boxes and change the order of assessment, when necessary, given the specific patient context	Provide option to click through guideline boxes to see criteria without selecting- ability to access all guideline boxes regardless of recommendation; display score of fulfilled criteria to show current suggestion for diagnosis, maintains control for clinician yet clear support
Visual display and aesthetics of user interface	Simple and minimalistic is good, not distracting	Webpage layout and not intuitive of how to navigate	Keep simple and minimalistic but also modern to help with engagement; provide navigation labels to allow for clinicians to explore how to navigate interface/use support
“Decision tree” style provision of decision support	Helped to organize structure of thought process and to keep track of decision making in line with the diagnostic criteria	Can feel like support is forcing a decision to be made with only yes or no option; does not fully reflect complexity of patient situations in CAMHS	Including a “I do not know” option so marked items will accumulate into box with summary of missing information to be able to efficiently acknowledge what is not known; use the diagnostic tree where you could look at symptoms falling above or below threshold for assessing the diagnosis
Accessibility of specific guideline criteria within guidelines	Can scroll through previously accessed guideline components	Limited guideline capacity to show all components until navigated through tree	Should provide option for going back/forward into the guideline specifics; if missing information should be able to look up where guideline provides support for criteria
Display of guideline support on user-interface next to patient information	Providing guideline support side by side with patient information could improve efficiency of clinical care; more quickly acknowledged what was needed because of accessibility	Decision support is available next to patient information but not connected so cannot interact with guideline and save any previously noted criteria met by patient	Future design of system should be able to have guideline support integrated with the patient to provide alerts relative to patients’ health information; ability to update criteria in between sessions and adjusts decision support provided to improve efficiency and overall coordination of services

the CDSS. The “decision tree” formatting might not optimally serve all clinicians, despite the formatting of the IDDEAS user interface and overall functionality design of the platform that could provide non-linear-based support for dynamic clinical care. In accordance with the need for improved coordination among services involved in CAMHS (4, 25), the IDDEAS project follows the Local Early Appropriate and Precise (LEAP) model (39). Our results reaffirm that as CAMHS in Norway depend on information coming from other services (i.e., educational and psychological counseling service, and the primary care provider), it is important to ensure the available patient information not only covers their current health status but also any previous care received from other social, school and health services (39, 40). This close collaboration provides the opportunity for customized guideline suggestions and availability of information about involved services, while also allowing early identification of risks. Early risk identification is a critical component of CAMHS to prevent the onset of mental health disorders (9).

Individualized Digital Decision Assist System development will continue to work toward full EHR integration, to keep collaborative efforts in CAMHS coordinated and ultimately to help to provide clinicians with accurate, efficient, and early clinical decision support through efficient and early identification of risks and provision of early intervention. With direct integration with the EHR, it will be possible to examine the potential for identifying previously addressed symptoms, while also identifying potential comorbidities by flagging relevant overlaps across multiple guidelines (5). An integrated CDSS potentially provides

specific, adapted suggestions relevant to the care of an individual patient, thus allowing the clinician to determine the extent to which they need to review other materials.

Maintaining this autonomy for the clinicians in CAMHS, allowing for them to be the decision-makers, is important for the acceptance and utility of future IDDEAS versions. As found by Kortteisto et al. (22) for the end-users to find a CDSS useful, they need to first trust it. As reported by Sutton et al. (13), diagnostic support based on patient data can be an advantage while also prove potentially harmful if users’ distrust what is provided by the CDSS. Graphical displays of statistics, access to scientific literature, and references to local EHR patient data patterns were all mentioned as examples of potential future design factors that help reassure clinicians of the CDSS trustworthiness while keeping the clinician as the main decision-maker (40).

The inclusion of support based on BUP-data is important to the clinicians as the end-users but is also important to service users (41). Service users in Norway want to be more involved in their care, including understanding the components of services received and sharing their data for the improvement of overall services (41). In general, clinicians want transparent presentations of EHR data that informs the decision support and recommendations provided for all stakeholders involved, making it more likely that stakeholders will trust and use the CDSS (21).

The results suggest that several attributes of the IDDEAS prototype should be addressed in the next version of IDDEAS (see Table 1). For the development of IDDEAS following UCD methods and an overall iterative approach (11, 30), we expect

continued identification of usability barriers and limitations to the design of IDDEAS, as seen commonly in other CDSS development studies (25, 38, 42, 43). For example, as similarly found by Baysari et al. (30) or Giordanengo et al. (43), it is important to design the system from the perspective of the end-user and finding overall improved perceptions of the system usability with the well-integrated EHR (44). The formative usability iterations underlying the design of IDDEAS promotes the ability to adjust as needed to meet the local coordinated CAMHS requirements, such as the provision of data-based recommendations from close integration with the patient EHR in the future. While there were several propositions for how to overcome the currently limiting attributes of the IDDEAS prototype, it will be important to also assess the identified barriers of the next prototype and provide a comparison to be able to understand the usability and potential utility of IDDEAS.

Strengths and limitations

There are multiple strengths and limitations of this study. First, there is a potential for bias. As participants were recruited by convenience sampling, this potentially attracted clinicians who may be already interested in innovation and potentially more comfortable interacting with technological solutions. While the focus of the study was to identify the CAMHS clinicians' perceptions of the IDDEAS prototype, it was a limitation that patients and their families were not involved as well (9, 25). Furthermore, as this is one component of a larger study within a multiple part project, the generalizability of our findings is limited. However, our sample size is in accordance with qualitative methodology standards and allows for un-saturated qualitative data and provides development information that will be used to design and execute larger scale IDDEAS usability studies. Additionally, the materials used in addition to the IDDEAS prototype (i.e., patient cases) also proved to have relative limitations, potentially deterring from the ability to assess the usability of the system attributes. However, as seen with other CDSS development studies following UCD designs, the patient cases provide a base for early on prototype testing and thus were intentionally designed to illicit important information for future research and help to identify current clinical needs.

Despite these limitations, there were strengths to the study as well. It was a strength to have the opportunity to work closely with the child and adolescent psychiatrists and psychologists who will ultimately be IDDEAS end-users. With the help and support from N-BUP we were able to recruit participants from multiple regions of Norway to help inform our continued development of IDDEAS. Furthermore, as participants were not provided with access to the IDDEAS platform prior to the usability evaluation and interviews, the study provides an authentic overview of the overall perceived usability, and specifically the learnability of IDDEAS at this stage. Our future research efforts will include iterations designated for service users testing and assessment of needs for optimal IDDEAS development. Despite qualitative research being highly dependent upon subjective interpretations and the relative competence of

both the researcher and interviewee (12, 13), the combination of inductive and deductive qualitative categories within the QCA was found to be in line with findings from other studies (13). The use of COREQ and SRSS allowed for us to minimize possible bias and ensure adherence to previously validated qualitative standards. The IDDEAS project's focus on formative usability assessments and prototype development allows for close collaboration with potential end-users in experimental settings, compensating for what could be deemed the limitation of qualitative interview inquiry. The mixed methods used in the usability evaluation will provide additional approaches to quantifying our findings, while also still allowing for efficiently identifying CAMHS clinicians' current needs and how IDDEAS can meet those needs.

Conclusion

Child and adolescent mental health services psychiatrists and psychologists shared the need for a completed IDDEAS clinical decision support system, especially if it is integrated within their clinical care processes; specifically, the electronic health record. Participants are eager to engage with the next phase, the dynamic high-fidelity prototype. Further usability assessments and identification of additional requirements for IDDEAS is necessary before feasibility testing and implementation. The findings from this study can help inform future IDDEAS development. A fully functioning, integrated version of IDDEAS has the potential to be an important contribution to support clinicians in the early identification of risks for mental disorders as well as full assessment and treatment of children and adolescents.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Regional Committee for Medical and Health Research Ethics, Southeast (REK Sør-Øst). The patients/participants provided their written informed consent to participate in this study.

Author contributions

CC was responsible for conducting the data collection, data analysis, and formulation of the manuscript. NS and BL were responsible for providing the guidance and feedback throughout the development of the original manuscript and contributing with edits to the manuscript. All authors contributed with their

Clausen et al.

10.3389/fpsy.2023.1033724

suggestions throughout the development of the final manuscript and read and approved of the final manuscript.

Funding

The IDDEAS project (269117) was funded and sponsored by the Norwegian Research Council (Norges Forskningsråd- NFR). The Norwegian Research Council will provide funding in line with HELSEVEL (Programme on Health, Care and Welfare Services Research), which promotes integrated patient and user pathways and research and innovation activities aimed toward improving the quality of expertise and efficiency in health care services. The Norwegian Research Council had no role in the design of the study, the collection, analysis and interpretation of data, and in writing the manuscript.

Acknowledgments

We would like to acknowledge the contributions from all clinicians who participated in this study and thank them for their time, and acknowledge N-BUP for their invaluable assistance with recruitment and overall support of the IDDEAS project research efforts.

References

- World Health Organization [WHO]. *Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*. (2022). Available online at: <https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide> (accessed March 5, 2022).
- Kessler R. Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry*. (2007) 20:359–64. doi: 10.1097/YCO.0b013e32816ebc8c
- McGorry P. Is this normal? Assessing mental health in young people. Adolescent health, focus. *Aust Fam Physician*. (2011) 40:94–7.
- Skokauskas N, Fung D, Flaherty L, Klitzing K, Puras D, Servili C, et al. Shaping the future of child and adolescent psychiatry. *Child Adolesc Psychiatry Ment Health*. (2019) 13:19. doi: 10.1186/s13034-019-0279-y
- Kessler R. Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO world mental health (WMH) survey initiative. *Int J Methods Psychiatr Res*. (2010) 19:4–22. doi: 10.1002/mpr.310
- McCoy B, Rickert M, Class Q, Larsson H, Lichtenstein P, D'Onofrio B. Mediators of the association between parental severe mental illness and offspring neurodevelopmental problems. *Ann Epidemiol*. (2014) 24:629–34. doi: 10.1016/j.annepidem.2014.05.010
- Rimvall M, Os J, Jeppesen P. Promoting a patient-centered, transdiagnostic approach to prevention of severe mental illness. *Eur Child Adolesc Psychiatry*. (2021) 30:823–4. doi: 10.1007/s00787-020-01563-y
- Shaligram D, Skokauskas N, Aragones E, Azeem M, Bala A, Bernstein B, et al. International perspective on integrated care models in child and adult mental health. *Int Rev Psychiatry*. (2022) 34:101–17. doi: 10.1080/09540261.2022.2059346
- Sampaio F, Feldman I, Lavelle T, Skokauskas N. The cost-effectiveness of treatments for attention deficit-hyperactivity disorder and autism spectrum disorder in children and adolescents: a systematic review. *Eur Child Adolesc Psychiatry*. (2021) 31:1655–70. doi: 10.1007/s00787-021-01748-z
- Clausen C, Leventhal B, Nytrø Ø, Kopusov R, Westbye O, Rost T, et al. Testing an individualized digital decision assist system for the diagnosis and management of mental and behavior disorders in children and adolescents. *BMC Med Inform Decis Mak*. (2020) 20:232. doi: 10.1186/s12911-020-01239-2
- Miller A, Moon B, Anders S, Walden R, Brown S, Montella D. Integrating computerized clinical decision support systems into clinical work: a meta-synthesis

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer FS declared a past co-authorship with the author NS to the handling editor.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsy.2023.1033724/full#supplementary-material>

- of qualitative research. *Int J Med Inform*. (2015) 84:1009–18. doi: 10.1016/j.ijmedinf.2015.09.005
- Schaaf J, Sedlmayr M, Sedlmayr B, Prokosch H, Storf H. Evaluation of a clinical decision support system for rare diseases: a qualitative study. *BMC Med Inform Decis Mak*. (2021) 21:65. doi: 10.1186/s12911-021-01435-8
- Sutton R, Pincock D, Baumgart D, Sadowski D, Fedorak R, Kroeeker K. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Med*. (2020) 3:17.
- Guenter D, Abouzahra M, Schabert I, Radhakrishnan A, Nair K, Orr S, et al. Design process and utilization of a novel clinical decision support system for neuropathic pain in primary care: mixed methods observational study. *JMIR Med Inform*. (2019) 7:e14141. doi: 10.2196/14141
- Dagliati A, Tibollo V, Sacchi L, Malovini A, Limongelli I, Gabetta M, et al. Big data as a driver for clinical decision support systems: a learning health systems perspective. *Front Digit Humanit*. (2018) 5:8. doi: 10.3389/fdigh.2018.00008
- Marco-Ruiz L, Pedrinaci C, Maldonado J, Panziera L, Chen R, Bellika J. Publication, discovery and interoperability of clinical decision support systems: a linked data approach. *J Biomed Inform*. (2016) 62:243–64. doi: 10.1016/j.jbi.2016.07.011
- Liberati E, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implement Sci*. (2017) 12:113. doi: 10.1186/s13012-017-0644-2
- Babione JN, Ocampo W, Haubrich S, Yang C, Zuk T, Kaufman J, et al. Human-centered design processes for clinical decision support: a pulmonary embolism case study. *Int J Med Inform*. (2020) 142:104196. doi: 10.1016/j.ijmedinf.2020.104196
- Berner E editor. *Clinical decision support systems: theory and practice, health informatics*. 3rd ed. Berlin: Springer International (2016).
- Kawamoto K, Houlihan C, Balas E, Lobach D. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. (2005) 330:765. doi: 10.1136/bmj.38398.500764.8F
- Chokshi S, Belli H, Troxel A, Blecker S, Blaum C, Testa P, et al. Designing for implementation: user-centered development and pilot testing of a behavioral economic-inspired electronic health record clinical decision support module. *Pilot Feasibility Stud*. (2019) 5:28. doi: 10.1186/s40814-019-0403-z

E-14 ► PAPER E USABILITY OF THE IDDEAS PROTOTYPE IN CHILD AND ADOLESCENT MENTAL HEALTH SERVICES

Clausen et al.

10.3389/fpsy.2023.1033724

22. Kortteisto T, Komulainen J, Mäkelä M, Kunnamo I, Kaila M. Clinical decision support must be useful, functional is not enough: a qualitative study of computer-based clinical decision support in primary care. *BMC Health Serv Res.* (2012) 12:349. doi: 10.1186/1472-6963-12-349
23. Nair K, Malaeekeh R, Schabort I, Taenzer P, Radhakrishnan A, Guenter D. A clinical decision support system for chronic pain management in primary care: usability testing and its relevance. *J Innov Health Inform.* (2015) 22:329–32. doi: 10.14236/ijhi.v22i3.149
24. Kopusov R, Fossum S, Frodl T, Nytrø Ø, Leventhal B, Sourander A, et al. Clinical decision support systems in child and adolescent psychiatry: a systematic review. *Eur Child Adolesc Psychiatry.* (2017) 26:1309–17.
25. HelseDirektoratet. *Aktivitetsdata for psykisk helsevern for barn og unge 2021*. Trondheim: HelseDirektoratet (2021).
26. Skokauskas N, Eckert M, Busch G, Andrade J, Park T, Guerrero A. Sustainable child and adolescent psychiatry. *Int Rev Psychiatry.* (2022) 34:97–100. doi: 10.1080/09540261.2022.2082163
27. NBUP. *En Interesseorganisasjon for Alle Avdelinger/Enheter Innen Psykisk Helsevern for Barn og Unge i Spesialhelsetjenesten.* (2023). Available online at: <https://nbup.no/> (accessed March 5, 2022).
28. Koochakpour K, Nytrø Ø, Westbye O, Leventhal B, Kopusov R, Bakken V, et al. Success factors of an early EHR system for child and adolescent mental health: lessons learned for future practice data-driven decision aids. *Stud Health Technol Inform.* (2022) 290:182–6. doi: 10.3233/SHIT220057
29. Clausen C, Leventhal B, Nytrø Ø, Kopusov R, Westbye O, Rost T, et al. Clinical decision support systems: an innovation approach to enhancing child and adolescent mental health services. *J Am Acad Child Adolesc Psychiatry.* (2021) 60:562–5. doi: 10.1016/j.jaac.2020.09.018
30. Baysari M, Duong M, Hooper P, Stockey-Bridge M, Awad S, Zheng W, et al. Supporting deprescribing in hospitalized patients: formative usability testing of a computerized decision support tool. *BMC Med Inform Decis Mak.* (2021) 21:116. doi: 10.1186/s12911-021-01484-z
31. Khajouei R, Eshfahani M, Jahani Y. Comparison of heuristic and cognitive walkthrough usability evaluation methods for evaluating health information systems. *J Am Med Inform Assoc.* (2017) 24:e55–60. doi: 10.1093/jamia/ocw100
32. Rouhbakhsh A, Badrfam R, Nejatiasafa A, Soori M, Sharafi SE, Etesam F, et al. Health care professionals' perception of stress during COVID-19 pandemic in Iran: a qualitative study. *Front Psychiatry.* (2022) 12:804637. doi: 10.3389/fpsy.2021.804637
33. Mayring P. Qualitative content analysis. *Forum Qual Soc Res.* (2000) 1:20.
34. O'Brien BC, Harris IB, Beckman T, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med.* (2014) 89:1245–51.
35. Kuckartz U. Qualitative text analysis: a systematic approach. In: Kaiser G, Presmeg N editors. *Compendium for early career researchers in mathematics education*. Berlin: Springer nature (2019). p. 181–97. doi: 10.007/978-3-030-15636-7_8
36. Mayring P. *Qualitative content analysis: theoretical foundation, basic procedures and software solution.* (2014). Available online at: <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173> (accessed May 1, 2022).
37. Genes N, Kim M, Thum F, Rivera L, Beato R, Song C, et al. Usability evaluation of a clinical decision support system for geriatric ED pain treatment. *Appl Clin Inform.* (2016) 7:128–42. doi: 10.4338/ACI-2015-08-RA-0108
38. Rost T, Clausen C, Nytrø Ø, Kopusov R, Leventhal B, Westbye O, et al. Local, early, and precise: designing a clinical decision support system for child and adolescent mental health services. *Front Psychiatry.* (2020) 11:564205. doi: 10.3389/fpsy.2020.564205
39. Fu X, Yang J, Liao X, Lin J, Peng Y, Shen Y, et al. Parents' attitudes toward and experience of non-suicidal self-injury in adolescents: a qualitative study. *Front Psychiatry.* (2020) 11:651. doi: 10.3389/fpsy.2020.00651
40. Bakken V, Kopusov R, Rost T, Clausen C, Skokauskas N, Nytrø Ø, et al. Attitudes of mental health service users toward storage and use of electronic health records. *Psychiatr Serv.* (2022) 73:1013–8. doi: 10.1176/appi.ps.202100477
41. Akhloufi H, Verhaegh S, Jaspers M, Melles D, Sijs H, Verbon A. A usability study to improve a clinical decision support system for the prescription of antibiotic drugs. *PLoS One.* (2019) 14:e0223073. doi: 10.1371/journal.pone.0223073
42. Ash J, Sittig D, McMullen C, Wright A, Bunce A, Mohan V, et al. Multiple perspectives on clinical decision support: a qualitative study of fifteen clinical and vendor organizations. *BMC Med Inform Decis Mak.* (2015) 15:35. doi: 10.1186/s12911-015-0156-4
43. Giordanengo A, Årsand E, Woldaregay A, Bradway M, Grottlund A, Hartvigsen G, et al. Design and prestudy assessment of a dashboard for presenting self-collected health data of patients with diabetes to clinicians: iterative approach and qualitative case study. *JMIR Diabetes.* (2019) 4:e14002. doi: 10.2196/14002
44. Kilsdonk E, Peute L, Kremer L, Jaspers M. Uncovering healthcare practitioners' information processing using the think-aloud method: from paper-based guideline to clinical decision support system. *Int J Med Inform.* (2016) 86:10–9. doi: 10.1016/j.ijmedinf.2015.11.011

ISBN 978-82-326-7600-2 (printed ver.)
ISBN 978-82-326-7599-9 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)



NTNU

Norwegian University of
Science and Technology