TAL TECH

DOCTORAL THESIS

Systematic AI Support for Psychiatry: A Framework on How to Implement Decision Support Systems

Markus Bertl

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2023

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
57/2023

Systematic AI Support for Psychiatry: A Framework on How to Implement Decision Support Systems

MARKUS BERTL

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Health Technologies

The dissertation was accepted for the defence of the degree of Doctor of Philosophy (Computer Science) on 26 October 2023

Supervisor: Prof. Dr. Peeter Ross, MD,
eMed Lab
Department of Health Technologies
School of Information Technologies
Tallinn University of Technology
Tallinn, Estonia

Co-supervisor: Prof. Dr. Dirk Draheim,
Information Systems Group
Department of Software Science
School of Information Technologies
Tallinn University of Technology
Tallinn, Estonia

Opponents: Prof. Dr. Martin Leucker,
Institute for Software Engineering and Programming Languages
University of Lübeck
Lübeck, Germany

Prof. Dr. Kerstin Denecke,
School of Engineering and Computer Science
Bern University of Applied Sciences
Bern, Switzerland

Defence of the thesis: 28 November 2023, Tallinn

Declaration:
I hereby declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Markus Bertl _____
signature

ISSN 2585-6898 (publication)
ISBN 978-9916-80-071-3 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-072-0 (PDF)
Printed by Koopia Niini & Rauam

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
57/2023

Tehisintellekti süstemaatiline kasutamine psühhiaatrias: otsustustoe rakendamist toetav raamistik

MARKUS BERTL

TAL TECH
KIRJASTUS

Contents

List of Publications

The present Ph.D. thesis is based on the following publications that are referred to in the text by Roman numbers [1].

I M. Bertl, J. Metsallik, and P. Ross. A Systematic Literature Review of AI-based Digital Decision Support Systems for post-traumatic Stress Disorder. *Frontiers in Psychiatry*, 13, 2022

II M. Bertl, P. Ross, and D. Draheim. A Survey on AI and Decision Support Systems in Psychiatry – Uncovering a Dilemma. *Expert Systems with Applications*, 202:117464, 2022

III M. Bertl, P. Ross, and D. Draheim. Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors. *Health Policy and Technology*, 2023

IV M. Bertl, K. J. I. Kankainen, G. Piho, D. Draheim, and P. Ross. Evaluation of Data Quality in the Estonia National Health Information System for Digital Decision Support. In *Proceedings of the 3rd International Health Data Workshop*. CEUR-WS, 2023

V M. Bertl, P. Ross, and D. Draheim. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *Database and Expert Systems Applications*, pages 104–111. Springer International Publishing, 2021

VI M. Bertl, M. Shahin, P. Ross, and D. Draheim. Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 826–833. Association for Computing Machinery, 2023

VII M. Bertl, N. Bignoumba, P. Ross, S. B. Yahia, and D. Draheim. Evaluation of Deep Learning-based Depression Detection using Medical Claims Data. *SSRN*, 2023[2]

[1] Please note that the order of the publications is not chronological according to publication date but represents the recommended reading order.

[2] Currently under review at Artificial Intelligence in Medicine

Author's Contributions to the Publications

I I was the main author of this publication. I defined the research problem and methodology, conducted the literature search with the co-authors, analysed the results, prepared some of the figures, and wrote the manuscript.

II I was the main author of this publication. I defined the research problem and methodology, conducted the literature search with the co-authors, analysed the results, prepared the figures, and wrote the manuscript.

III I was the main author of this publication. I defined the research problem and methodology, conducted the literature search, analysed the results, prepared the figures, and wrote the manuscript.

IV I contributed equally with the second author of this publication by creating the research design and writing the introduction and conclusion sections. Additionally, I did the analysis of the primary use of health data and wrote the corresponding sections in the methods, results, and discussion chapters.

V I was the main author of this publication. I defined the research problem and methodology, conducted the literature search, developed the machine learning scripts, analysed the results, prepared the figures, and wrote the manuscript.

VI I was the main author of this publication. I defined the research problem and methodology, developed the ARM JupyterLab scripts, analysed the results, and wrote the manuscript.

VII I contributed equally with the second author. I defined the research problem, worked on the data collection and data pre-processing, contributed medical domain knowledge to the algorithm development, and co-authored the manuscript. The research methodology was developed jointly with the second author.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ARM	Association Rule Mining
AUC	Area Under the Curve
AUPRC	Area Under Precision-Recall Curve
CNN	Convolutional Neural Network
DDSS	Digital Decision Support System
DHP	Digital Health Platform
DICOM	Digital Imaging and Communications in Medicine
EHIF	Estonian Health Insurance Fund
EHIS	Estonian Health Information System
EHR	Electronic Health Record
EMR	Electronic Medical Record
FNN	Feed Forward Neural Network
GP	General Practitioner
GRU	Gated Recurrent Unit
HL7 CDA	Health Level 7 Clinical Document Architecture
ICD-10	International Classification of Diseases, 10th Revision
LOINC	Logical Observation Identifiers Names and Codes
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
PTSD	Post-traumatic Stress Disorder
QUALY	Quality-Adjusted Life Year
RQ	Research Question
SLR	Systematic Literature Review
SNOMED-CT	Systemized Nomenclature of Medicine – Clinical Terms

Terms & Definitions

Artificial intelligence	is the study of how to produce computer programs that have some of the qualities of the human mind. For example, the ability to understand language, recognize pictures, solve problems, and learn [57].
Clinical vs. medical	There are many definitions for the words clinical and medical. For this research, we see 'clinical' as everything that is practiced on the patient in clinical conditions, i.e., diagnosis, treatment, or rehabilitation. 'Medical' refers to the much wider domain that, besides clinical activities, includes other activities that have a connection with human health, for instance, biomedicine, genetics, or healthcare technology.
Digital Decision Support System (DDSS)	is a computer-based system that brings together information from a variety of sources, assists in the organization and analysis of information and facilitates the evaluation of assumptions underlying the use of specific models [67]. Sometimes also referred to as Clinical Decision Support System (CDSS).
Framework	is a system of rules, ideas, or beliefs used to plan or decide something [72].
Incidence vs. prevalence	Incidence is the probability of occurrence of a medical condition in a population during a specific time period while prevalence is the proportion of a population affected by a medical condition at a specific time.
International Classification of Diseases, 10^{th} Revision (ICD-10)	is a classification and coding tool maintained by the WHO. It contains codes for diseases, disease descriptions, and symptoms.
The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)	is a commonly used taxonomic and diagnostic tool for mental disorders published by the American Psychiatric Association [6]. Compared to the ICD, the DSM is only focused on psychiatric disorders.
Symptoms	are single representations of a disease. Several symptoms form the syndrome. One symptom could be the same for different diseases. Though, a bundle of symptoms refers to a syndrome that is specific in cases where all necessary symptoms are present.
Disease vs. disorder	Disease is an objective, measurable pathological process or finding of a person which is described according to specific symptoms and pathomorphologies in defined taxonomies. Disorder refers to irregularities, disturbances, or interruptions in a person's health condition, somatic or psychological, which are observable but difficult to measure objectively. In mental health, the term disorder is often preferred and also used by ICD-10 and DSM-5.

Summary

This research summary is structured as follows. In Sect. 1, I establish the motivation, provide the problem statement behind this work, give an overview of the state of the art and related work (based on publications [I], [II] and [III]), and state the research questions (RQs) addressed. In Sect. 2, I provide an overview of the research methodologies used, leading to the contributions made by this work, which are described in Sect. 3 (based on publications [I], [II] and [III]) and applied in Sect. 4 (based on publications [IV], [V], [VI] and [VII]). Related work, contributions, limitations, and further research are discussed in Sect. 5, and the conclusion can be found in Sect. 6.

1 Introduction

1.1 Research Relevance and Medical Background

As the leading cause of years lived with disabilities, undiagnosed psychiatric disorders not only induce avoidable suffering [84, 81], but also impact society at large [41]. In 2010, diseases of the brain accounted for €461 billion in healthcare costs in Europe [36]. For undiagnosed depression alone, the quality-adjusted life years (QUALYs) lost amount to $9,950 per affected citizen in the US [83]. Diseases and disorders of the brain are, with an annual prevalence of about 38%, very common [84]. Hence accurate diagnosis and effective treatment of those diseases have a high impact on our global wellbeing. Additionally, prevalence has not decreased over the years, showing that current healthcare systems are not addressing this issue adequately [81].

The following paragraph describes what the clinical process for diagnosing psychiatric disorders looks like in most European countries. The general practitioner (GP), also called family physician, is often the first point of contact for people who are feeling unwell. The GP examines the patient and makes an initial diagnosis. If a psychiatric disorder is suspected, the GP typically refers the patient to a specialist like a psychologist or psychiatrist for confirmation of the diagnosis and initiation of treatment.

In parallel, the documentation process is started. For the easier handling of diagnosis data in IT systems, diagnoses are encoded based on disease classification systems like the International Classification of Diseases, 10th Revision (ICD-10)[3]. An ICD-10 code consists of alphanumeric characters which clearly identify a disease. The ICD has a hierarchical tree structure where each level in the tree adds additional information or a specification to the disease (category). As an example, the code F33.2 can be parsed from left to right where F codes all mental and behavioral disorders, F3 Mood [affective] disorders, F33 Recurrent depressive disorder, and F33.3 "Recurrent depressive disorder, current episode severe without psychotic symptoms". The coding responsibility generally lies with the physician but is often outsourced to specialists or automatized. In most countries with a public health insurance system, those coded diseases or medical interventions are then digitally transmitted to the insurance provider for billing. Additionally, medical professionals create documentation that is saved in the physicians' electronic medical records (EMR). These data consist of structured data, like vital signs, as well as unstructured data, like free text or medical images. Often, documentation in an unstructured format is preferred because it is easier and faster for the healthcare professional compared to filling in structured forms. Depending on the eHealth maturity of the country, EMR data can then be pushed into a central electronic health record (EHR) of a national digital health platform (DHP). There, data are mostly saved based on international standards like HL7

[3] https://icd.who.int/browse10/

CDA[4], DICOM[5], LOINC[6], and SNOMED-CT[7].

We often perceive the way humans work as the gold standard. In the medical domain, especially in psychiatry, some evidence shows that this "gold standard" is severely flawed. Patients are not treated according to medical guidelines [56], and diagnostic accuracy is generally low [1, 3, 38]. Mitchell et al. reported that only 52.7% of people with depression are correctly diagnosed [60]. General practitioners (GPs), who are often the first point of contact for patients entering the healthcare system, only have a 50.1% (95% CI: 41.3 to 59.0) sensitivity for diagnosing depression [60]. As another example, the number of wrongly diagnosed people with bipolar disorder is about 69% [74]. Patients stay with no diagnosis or the wrong diagnosis for approximately 5 to 7.5 years [62].

Symptoms that could indicate psychiatric disorders are typically very general and difficult to distinguish from other physical illnesses like sleep disorders, headaches [52], or pain in the musculoskeletal system. Moreover, the symptom patterns sometimes span over a large number of doctors' visits. This makes symptoms difficult to notice. While other medical conditions can be diagnosed using lab tests, medical imaging, or other quick and specific diagnostic tools, biomarkers to diagnose psychiatric disorders are still far away [51]. The recommended way to diagnose psychiatric illnesses is based on questionnaires and assessment scales [4]. However, in order to administer them, a trained medical professional who takes between 10 and 30 minutes [9, 39, 61] is needed. To put this into perspective, an average GP visit lasts approximately 11 minutes [65], even visits to specialist doctors, like psychiatrists, only last between 19 and 21 minutes [70]. Additionally, the current "state of the art" diagnostic methods in psychiatry date back to the early 1960s [9, 39]. However, the exact definition of mental disorders, especially the distinction between normality and psychopathology is still a subject of debate under medical professionals [76]. Because of the continuously changing understanding of what is considered to be "pathological", together with the development of society in general, the content of those questionnaires might not reflect the current state of the art anymore. A prime example is homosexuality, which was considered a psychiatric disease by the American Psychiatric Association at the time some of the currently used diagnostic tools were created [25]. Those challenges influence the fact that the clinical method of diagnosis, where decision-makers have no universally agreed processes or taxonomies to compile and evaluate findings for diagnosis and treatment [23], is frequently used. Nevertheless, the actuarial method of diagnosis, which eliminates the human judge and is solely based on empirically established relationships between data and the decision to diagnose, should be preferred [58, 73, 23]. The situation of the high number of wrong or missing psychiatric diagnoses could be improved using the actuarial method [73].

1.2 Artificial Intelligence & Decision Support – Status Quo

A prime example of actuarial thinking is represented through Artificial Intelligence (AI). AI research was founded as a research field in the 1950s and was largely based on the ideas of Alan Turing and John von Neumann [28]. Since then, the domain has survived several so-called AI winters, where, after a period of fast-growing enthusiasm, people became discouraged in the technology due to the lack of practical results [37]. From 2009 until now, AI has been discussed with an increasingly optimistic tenor [28]. More and more AI successes have been triggered through the increase in affordable computing power

[4] http://www.hl7.org/
[5] https://www.dicomstandard.org/
[6] https://loinc.org/
[7] https://www.snomed.org/

and data storage capabilities, but also because of the introduction of deep learning [48]. Nowadays, AI plays a key role in many domains by boosting efficiency through process automation and data-driven decision-making, and by uncovering previously hidden patterns and facts from data more accurately [5]. According to a global McKinsey survey, AI adoption has more than doubled since 2017 [20]. The companies surveyed by McKinsey claim that AI accounted for an increase of more than 5% in their earnings before interest taxes (EBIT) [20]. Moreover, AI also reached the consumer space. The recent hype around Large Language Models, especially Generative Pre-Trained Transformers like ChatGPT, is just one of many, some may allege, success stories of AI. AI in health care is wrapped in so-called Digital Decision Support Systems (DDSSs). DDSSs are AI-based systems that aid clinicians in their decision-making processes to improve healthcare delivery [78]. In the early 1970s, rule-based expert systems like INTERNIST-I were developed as the first DDSSs [55]. However, the uptake of AI in medicine has been slower compared to other domains, largely because digital data availability has only started to increase in recent years [5]. One groundbreaking milestone was IBM's Watson winning the quiz show Jeopardy! in 2011. Based on Watson's backward reasoning and natural language processing capabilities, several applications for AI in health care have been envisioned [29, 59], and some even implemented [7, 10]. Besides Watson, researchers proposed DDSS technology based on other AI algorithms for many domains like radiology [46], dermatology [26], or internal medicine [82]. It was argued as early as in 1987 that AI would take over the intellectual function of physicians [68]. In 2016, the announcement was made that it no longer makes sense to train radiologists [19]. However, as of 2023, these predictions have not come true. AI in health care still seems to be overpromised and underdelivered [77, 8, 79]. Furthermore, for the domain of psychiatry, based on the numbers presented in Sect. 1.1, AI seems to have not yet brought about a large improvement in the situation of patients or physicians. To further investigate the current state of the art in research on DDSSs in psychiatry, Sect. 1.3 presents a systematic overview of the current state of AI research in psychiatry.

1.3 State of the Art of DDSSs in Psychiatry – Literature Overview

This section presents the summary of two systematic literature reviews (SLRs) carried out by the author and published in publication [I] and publication [II]. For this, 585 research articles about DDSS in Psychiatry were analysed. Publication [I] presents an SLR about DDSSs for post-traumatic stress disorder (PTSD) to investigate the state of the art in DDSSs for a specific psychiatric disorder; publication [II] presents an SLR about DDSSs in psychiatry in general. The main outcomes relevant to this work were the following:

- DDSS prototypes in the research articles have generally low maturity levels (as defined in Sect. 3.7) and do not demonstrate clinical value in most cases.
- The DDSS prototypes therefore mainly focus on decision algorithm development. Hence, the DDSS evaluation also focuses on algorithmic accuracy metrics (e.g., accuracy, AUC, APURC).
- Sample sizes of training and testing data in current DDSS research are often low (median of 151.5 records).

Based on those findings, two major research gaps were identified. One in terms of data and decision technology. Figure 1 presents the sample sizes and corresponding accuracy of the decision technologies included in the literature reviews carried out. It can be seen that studies with high sample sizes are largely missing (six studies based on 5972, 11,540,

14,929, 45,388, 89,840, and 89,840 samples have been removed as outliers from the plot for better readability). However, a high sample size is vital to assess whether the AI algorithms' performance is one reason for the low maturity rates found in the literature review. The low sample sizes on which algorithms have been trained and evaluated also make the external validity of the claimed success rates of many DDSS algorithms questionable because of potential overfitting or selection bias in the data [40]. Evaluation of decision technology on large, real-world, realistic datasets is needed to be able to indicate whether AI technologies deliver adequately high accuracy metrics to improve the situation in psychiatry.

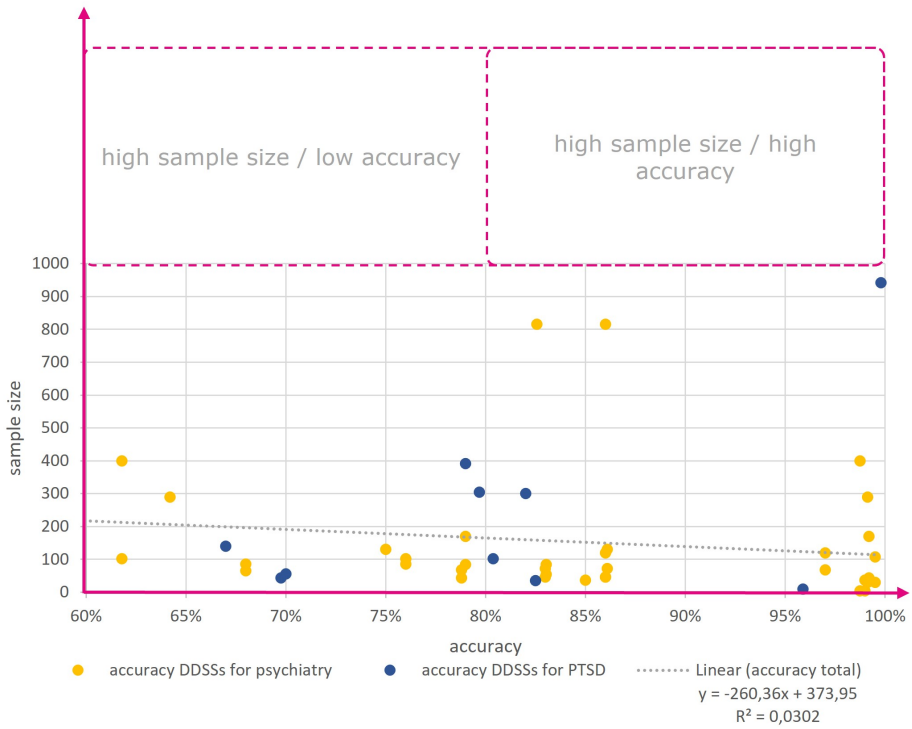


Figure 1: Scatter Plot DDSS Sample Size vs. Accuracy

The second research gap is the lack of holistic research, which takes into consideration not only a specific area like the algorithmic part of the decision technology, but also additional factors like real-world data availability, medical domain knowledge, clinical processes, user interaction, or validation of the whole system. The analysed studies rarely applied a systematic approach that takes into consideration all or several of the abovementioned factors. This narrow focus leads to the development of DDSS fragments, which are created in isolation and under "textbook" conditions. Such artifacts often do not reach higher maturity levels because they start to fail under real-world conditions since the artifact is not connected to the medical or business process. The lack of a holistic approach, ultimately, could be one reason for the low adoption rate of DDSS research in medical practice.

1.4 Research Questions

The aim of this work is to improve the development process of DDSSs in medical practice. Therefore, this thesis fills two major research gaps: first, the lack of a systematic approach and well-described knowledge regarding the theoretical and technological aspects of DDSSs, which leads to the second, low adoption rates. The primary research question (RQ) of this work is "*How can the design and development of AI-based DDSSs in psychiatry be improved?*". To provide a clear scope and decrease complexity, this RQ was split into the three sub-research questions stated below. Table 1 presents the mapping of each sub-RQ to the corresponding publications that contribute to answering it.

- *sub-RQ1*: What are the current obstacles hindering the adoption of AI-based DDSSs in psychiatry?
- *sub-RQ2*: How can DDSSs bring value to clinicians?
- *sub-RQ3*: Which AI approaches are best suited, comparatively, to diverse scenarios of DDSSs implementation in psychiatry?

Table 1: Mapping of associated RQs and publications

Research Question	*Publications*
RQ	[I], [II], [III], [IV], [V], [VI], [VII]
sub-RQ1	[I], [II], [III], [IV]
sub-RQ2	[I], [II], [III]
sub-RQ3	[V], [VI], [VII]

2 Research Methodology

This summary is composed based on 7 original, peer-reviewed research articles (4 journal articles and 3 conference papers). Publications [I] and [II] contribute through analysis of the state of the art of DDSS in psychiatry and propose a novel artifact designed as a conceptual framework to improve the design and development of DDSSs and ultimately raise adoption rates. The overall research process is further described in Sect. 2.1, and the framework creation in Sect. 2.2. Publication [III] shows how the framework can be applied and then contributes by critically evaluating the framework and adding practical insights through a focus group interview. The evaluation method is further described in Sect. 2.3. Publication [IV] evaluates the data quality of data sources for AI-based DDSSs. Lastly, Publications [V], [VI] and [VII] propose and evaluate DDSS decision technology prototypes using large, real-world data from the Estonian Health Insurance Fund (EHIF). Publication [V] contributes by showing a machine learning approach, [VI] a rule-based approach, and [VII] a deep learning-based approach for DDSS decision technology. The method of prototype development is described in more detail in Sect. 2.4. In order to strengthen the internal validity of the presented results, multiple sources of evidence were used to triangulate the conclusions of this research [24]:

- **Literature** to assess the current state of the art.
- **Subject matter experts** and the authors' own experience to better understand the practical problem and verify potential solutions.

- **Medical data** e.g., from electronic health records (EHR) from the Estonian Health Information System (EHIS) and databases like the data warehouse of the EHIF, to design and evaluate systems.

A graphical overview of how the publications are linked to the overall outcomes of this research is shown in Figure 2.

2.1 Research Process

For the work presented in this thesis, the design science research paradigm proposed by Hevner et al. was utilized [43]. Design science is a systematic methodology for developing novel artifacts (e.g., technology or frameworks) that cope with real-world problems. The designed artifact needs to solve a specific problem that is rigorously defined, formally represented, coherent and internally consistent, and comprehensively evaluated [42]. Design science has been widely accepted as an information systems research method [42, 47]. Based on the above-mentioned three pillars, I developed the design science research process shown in Figure 2, resulting in the framework described in Sect. 3 and the prototypes described in Sect. 4.

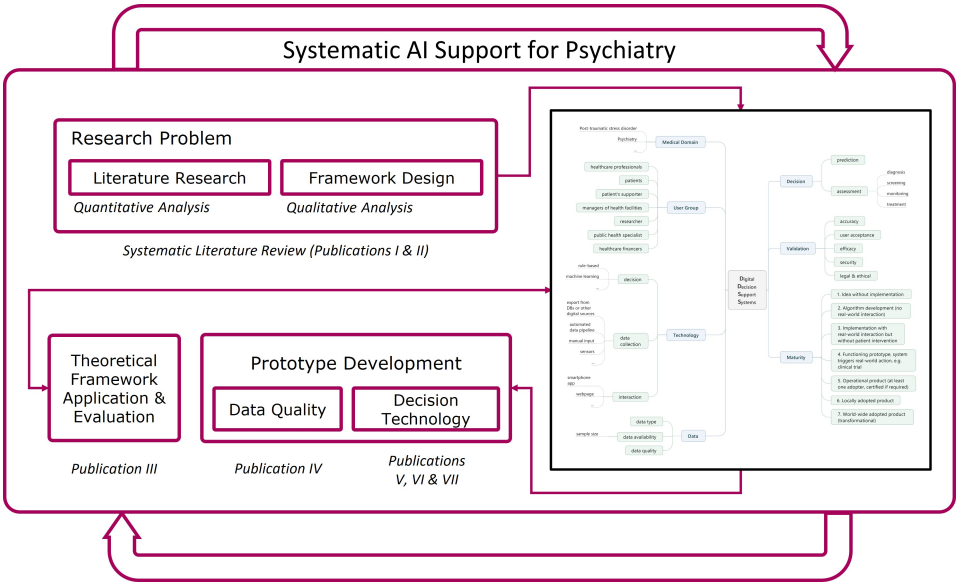


Figure 2: Research process. The figure shows the research process of the work together with the corresponding publications (in italics)

2.2 Quantitative and Qualitative Analysis – Framework Design

We used a systematic literature review based on the guidelines from Kitchenham & Charters [49] and the PRISMA guidelines [53] to quantitatively explore the state of the art of DDSSs in psychiatry. Based on those results, we further used thematic analysis [22] to aggregate the results into a conceptual framework for DDSSs (see Figure 3).

2.3 Evaluation

The evaluation of the framework has been carried out on the basis of two main methods:

- We used a *focus group* [64] with nine experts on DDSSs to validate the literature-derived framework. Additionally, these experts were able to bring in domain knowledge and practical experience to make sure that the framework is not only based on scientific literature but also serves as a tool that can be applied in real-world DDSS development and evaluation scenarios. Here, the triangulation based on the use of evidence like scientific literature, domain expertise obtained through the focus group interview, and the expertise of the authors strengthens the internal validity of the proposed artifact.

- Additionally, a *scenario-based evaluation* [43] was carried out to demonstrate the utility of the proposed artifact. Based on this scenario, each dimension of the framework was explained to demonstrate its usefulness. The scenario is described in Sect. 3.

2.4 Prototype Development

As part of this work, DDSS prototypes were developed (see Sect. 4). We investigated the three possible decision technologies: traditional machine learning-based [V], rule-based [VI], and deep learning-based [VII]. Most DDSS research has only been evaluated on small datasets [I], [II]. To obtain trustworthy results on how well these three decision technologies perform, they were applied to real-world data. The machine learning and rule-based prototypes were developed based on diagnosis, diagnosis date, and demographic data (birth year, sex) from the EHIF's data warehouse. We obtained anonymised information from 60,115 adults (18 years or older) with a total of 904,821 ICD-10 coded diagnoses between 2018 and 2019. The data consist of all publicly insured people in Estonia with a depression diagnosis, either single episode (F32) or recurrent (F33), and an equally-sized random sample of people with other psychiatric disorders. The percentage of insured people in Estonia is above 93.63% [27], so we are confident that our dataset is representative of the whole Estonian population. Since the deep learning-based prototype was developed last, more recent data were included. The used dataset consists of 812,853 patients (all people with a psychiatric disorder as well as a random sample without a psychiatric disorder) with a total of 26,973,943 diagnoses between 2018 and 2022. The Research Ethics Committee of the National Institute for Health Development (TAIEK[8]) approved the research design and data usage for the prototype development (Decision No. 1148).

3 The DDSS Framework

As described in Sect. 1.3, artifacts that do not take into consideration a broader perspective of their domain have a low chance of reaching high maturity levels because they fail under real-world conditions. The developed DDSS framework serves as a boilerplate for a systematic approach to DDSS development and analysis, or in other words, a systematic AI approach to psychiatry. A detailed explanation of the framework can be found in publication [III]. In the following subsections, the framework dimensions are explained based on a specific DDSS scenario. Currently, doctors are struggling to diagnose psychiatric disorders in a timely manner. As mentioned in Sect. 1.1, GPs display low sensitivity when it comes to diagnosing mental disorders [60][74]. People go undiagnosed for 5 to 7.5 years on average [62]. To diagnose people faster and more accurately, we propose a

[8] Tervise Arengu Instituudi inimuuringute eetikakomitee

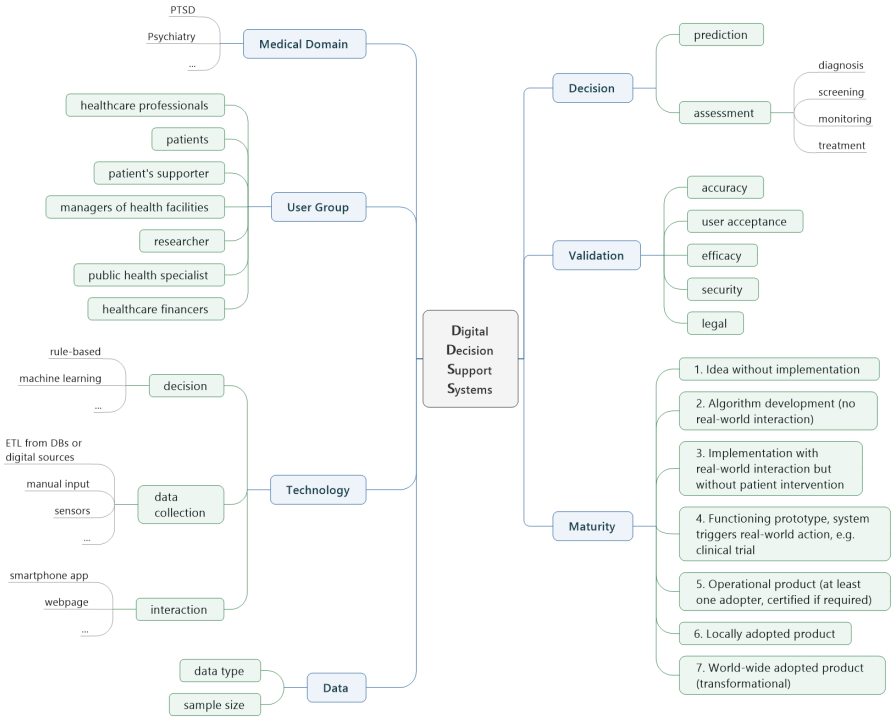


Figure 3: Developed DDSS Framework for Systematic AI Support from [III]

DDSS that utilizes previous diagnoses of patients to give an indication as to whether this patient might have an underlying, currently undiagnosed psychiatric disorder. An example patient journey could look like this: The patient comes to a family physician complaining of headaches. The patient gets pain medication prescribed, and an ICD-10 coded diagnosis like R51.9 (Headache, unspecified) is entered into the physician's IT system for documentation and billing purposes. A month later, the same patient comes to the GP with constant, mild back pain. The patient gets a prescription for an ointment and the diagnosis M54.5 low back pain. When the physician enters this diagnosis, the proposed DDSS, which works in the background and screens all past medical events entered into the patient's EHR, understands that these symptoms the patient described could be consistent with an overlooked psychiatric disorder. It raises an alert in the physician's system to make them aware of a potential underlying psychiatric disorder and shows the disease patterns that led to this conclusion. The physician can now validate the suggestion and, if needed, refer the patient to a specialist to confirm and treat the psychiatric disorder.

Please note that this scenario description has been simplified to be concise for this summary. Realistically, the disease pattern leading to the DDSS alert would be much more complex and span over more visits, potentially even at different healthcare providers. However, the provided example is sufficient to demonstrate the main idea and highlight how to use the proposed framework for systematic AI support (see Figure 3). Each of the following subsections now applies one dimension of the framework to the scenario.

3.1 Data

The data dimension describes the data needed for a DDSS to function. Important factors are the type of data that should be used and the quantity needed in the right quality for training, evaluation and results output. Since research on low maturity levels typically uses public or other easily available datasets, it is crucial to investigate whether these data are also available in the same structure, quality and quantity during the clinical process in which the DDSS should be applied. For the described scenario, the usage of medical claims data has been chosen. This type of data is easily available in large quantities and sufficient quality in most countries, and covers a large part of the population [IV]. Medical claims data are typically structured and consist of demographic information (ID, sex, birth date, etc.) as well as one or more coded diagnoses or interventions. In our scenario, the coding system ICD-10 is used.

3.2 Technology

The technology dimension describes, based on three categories, how a DDSS is implemented:

- *Decision technology* is the algorithm that powers the decision-making of the DDSS. Based on the findings of publication [I] and publication [II], this is the most researched component of DDSSs. Prototypes for the proposed scenario of the decision technologies – machine learning [V], rule-based [VI], and deep learning [VII] – are shown and explained in Sect. 4. Additionally, hybrid approaches with a mix of decision technologies are possible.

- *Interaction technology* describes how the system interacts with different user groups and/or the clinical process. Examples could be application programming interfaces (APIs), graphical user interfaces (GUIs), or sensory input from videos or speech. In the chosen scenario, the interaction technology could be the available IT system of the GP, the platform of the health insurance fund or the patient summary tab of

the national EHR system. When the physician opens the patient's data, the DDSS can automatically raise a flag in the case of an assumed undiagnosed psychiatric disorder to make the doctor aware of it. This process has already been positively evaluated for the Estonian Drug-Drug Interaction System [50].

- *Data collection technology* defines how the data the DDSS needs for the decision technology are gathered. Since medical claims data are generated from the GP's IT system and transferred to the EHIF and NHIS, for billing and documentation purposes, data collection can be done directly from their data storage using Extract Transform Load (ETL) technologies. No additional data capturing is needed.

3.3 User Group

This dimension deals with the users involved with the DDSS and how the DDSS should be infused into their work processes to provide maximum value. In other words, this dimension should analyse what a DDSS needs to do, at which step of the clinical process and in which way in order to support medical professionals. The suggested user group in this scenario is GPs, as they are most often the first point of contact for patients entering the healthcare system. Therefore, most patients can be reached like this. Since the proposed DDSS uses only data generated during the diagnosis process, no major changes to the clinical process for data capturing are required. Furthermore, in the case of a positive alert from the DDSS, no major process changes are required. The GP can consider the DDSS output as just another factor in the decision-making process.

3.4 Medical Domain

The medical domain dimension deals with the medical background knowledge needed in order to provide a functioning DDSS that brings real value to medical professionals. Examples include the decision a system should support in order to improve the process as well as the way knowledge is transferred to the DDSS. In other words, what the DDSS learns and how. For the proposed DDSS scenario, the medical background knowledge needed is in the areas of psychology, psychiatry, and family medicine. As shown in Sect. 1.1, the medical background investigation shows that the use case contributes to solving an actual problem in the domain. The learning of the DDSS is based on past medical diagnoses or interventions noted in medical claims data, which is a clinically valid data source. If the process of creation of these data is investigated, several potential reasons for bias can be found. For example, since the entered diagnosis is used for billing, a monetary incentive to put specific diagnoses over others is created. These potential biases introduced by the clinical process need to be understood, closely monitored, and mitigated.

3.5 Decision

The decision dimension deals with the output of the DDSS and how this output is used. In the described scenario, the decision would be an assessment of the current patient status. Since the goal is to detect psychiatric disorders quickly, on a large number of patients, and as early as possible, it can be classified as a screening use case.

3.6 Validation

The validation dimension deals with the measurements of success of DDSSs. In other words, how is the DDSS evaluated to ensure that it works and brings benefit. There are many types of validation, such as accuracy, user acceptance, efficacy, compliance, security, or legal validation. For the described scenario, the decision technology was evaluated using algorithmic accuracy, with an additional focus on practicality and user acceptance.

3.7 Maturity

To assess the DDSS's development status and the transition of DDSS research into clinical practice, a novel maturity scale has been developed. It is based on technology readiness levels from NASA[9] but has been adapted by us in publication [I] to better suit DDSSs in health care (see Table 2). Typically, the lower the maturity level, the fewer dimensions of

Table 2: DDSS Maturity Levels

Level	Description
1	Idea without implementation
2	Implementation without real-world interaction (algorithm development)
3	Implementation with real-world interaction but without patient intervention (no real intervention on a patient takes part based on the output of the DDSS)
4	Fully functioning prototype, system triggers real-world action (e.g. clinical trial)
5	Operational product (at least one adopter, certified if required)
6	Locally adopted product
7	Globally adopted product (transformational)

the proposed DDSS framework are considered in DDSS research. To increase the maturity level of a proposed artifact, a broader and more detailed analysis of our framework dimensions should be carried out by DDSS researchers to ensure that the necessary information for a successful system adoption is taken into consideration. Our proposed DDSS prototype for the mentioned scenario currently has a maturity level of 2. The clinical trial to reach level 4 is in planning.

4 Framework Application – Prototype Design

The following section describes the prototypes of DDSS decision technologies that have been implemented and evaluated. Decision technology of DDSSs can be implemented using three main approaches (or a combination of those): Machine learning as described in Sect. 4.1, rule-based as described in Sect. 4.2, and deep learning-based as described in Sect. 4.3.

4.1 Machine Learning

This subsection summarizes the results of publication [V]. Here, the results of the publication are used to evaluate traditional machine learning (ML) methods as decision technology for the proposed DDSS scenario. In general terms, ML has been defined by [32] as "the systematic study of algorithms and systems that improve their knowledge or performance with experience". Please note that deep learning algorithms, as a subset of ML, also meet this definition but are excluded in this section and investigated separately in Sect. 4.3. To effectively evaluate the most common traditional ML algorithms for classification tasks, AutoAI/AutoML was used. AutoAI uses technologies like Bayesian optimization, meta-learning, and ensemble construction to automate the ML life-cycle end to end (e.g., data preparation, feature engineering, model selection, pipeline optimization, hyperparameter optimization) [31]. Ideally, AutoAI should therefore be able to take a given dataset, analyse it, automatically transform and engineer features, test the classifiers, optimize the hyper-parameters, and return the best ML model for the downstream task. Based on

[9] `https://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level`

this technology, we used the dataset described in Sect. 2.4 to evaluate the accuracy of 15 classifiers (with 14 feature engineering techniques and 4 data pre-processing methods [30]) for predicting psychiatric disorders. We did not apply any data pre-processing or manual feature engineering before feeding the data to the AutoML library Auto-Sklearn. Detailed information on Auto-Sklearn can be found at [30, 31]. The average accuracy of all runs of the best classifier ensemble selected by Auto-Sklearn for the diseases and disease category shown in Table 3 is 0.6, 95% CI [0.596, 0.604], with a F_1-score of 0.58, a precision of 0.61, and a recall of 0.56. According to our literature survey, the average ac-

Table 3: Auto-Sklearn classifiers – average performance

Disease	Precision	Recall	F_1-Score	Accuracy	Number of Test Data Records
F32	0.60	0.56	0.58	0.59	4331
F33	0.63	0.57	0.6	0.61	4325
F43	0.59	0.63	0.61	0.59	1195
F	0.63	0.47	0.53	0.60	17194

curacy of DDSS algorithms in psychiatry is about 82.8% [16], i.e. much higher than in our results. We assume that this is because traditional ML algorithms often struggle to deliver satisfactory results when applied to complex and large datasets due to their limited capacity to capture intricate relationships within the data and handle high-dimensional feature spaces. Moreover, the inherent heterogeneity and noise in healthcare data can further exacerbate these limitations.

4.2 Rule-based

This subsection summarizes the results of publication [VI], which describes how Association Rule Mining (ARM) [44] can be used to create rule candidates for the knowledge base of a rule-based DDSS. ARM finds associations and correlations throughout large sets of data and provides information in the form of 'if-then' statements [71]. For this research, the Apriori algorithm [2] was used to mine association rules based on the dataset described in Sect. 2.4 to find out which disease codes often co-occur. Table 4 presents the association rules that have a specific psychiatric diagnosis as a consequent. Table 5 translates the ICD-10 codes of Table 4 to the corresponding textual description. We also used clustering based on the hierarchical structure of the ICD-10 codes to see if there are differences in certain granularity levels. Indeed, a higher number of interesting association rules were found by clustering all psychiatric disorders into one group. The table with detailed results can be found at [VI]. This symbolic AI approach of building a DDSS based on rules found using ARM requires only limited computing power, and decisions based on rules are fully transparent. This full transparency allows for an easy impact analysis on the influence of certain rules to an output. While re-training a machine or deep learning model can lead to unpredictable outcomes, the impact of fine tuning or changing the rules on the system can easily be assessed. However, rules need to be evaluated and selected manually from all the rule candidates found. This increases human involvement in the DDSS knowledge base creation, which leads to high costs and time for knowledge acquisition and maintenance. Furthermore, rules based on this ARM approach do not take into consideration the time dimension, sequence, or frequency of patients' visits. Therefore, the decision technology has a limited ability to find deeply hidden disease or behavioural patterns in data.

Table 4: Association rules of ICD-10 codes without F-clustering

#	antecedents	consequents	ant. sup.	con. sup.	support	confidence	lift	leverage	conviction
1	(I11)	(F51)	0.1602	0.0586	0.0168	0.1051	1.7935	0.0074	1.0519
2	(M17)	(F33)	0.0615	0.1861	0.0157	0.2553	1.3718	0.0043	1.0929
3	(K21)	(F33)	0.0639	0.1861	0.0159	0.2490	1.3379	0.0040	1.0837
4	(G47)	(F33)	0.0778	0.1861	0.0193	0.2479	1.3321	0.0048	1.0822
5	(R10)	(F41)	0.0822	0.1976	0.0215	0.2621	1.3264	0.0053	1.0873
6	(K21)	(F41)	0.0639	0.1976	0.0167	0.2604	1.3178	0.0040	1.0849
7	(G47)	(F41)	0.0778	0.1976	0.0197	0.2537	1.2839	0.0044	1.0751
8	(N30)	(F41)	0.0746	0.1976	0.0187	0.2506	1.2682	0.0039	1.0706
9	(G47)	(F32)	0.0778	0.2110	0.0206	0.2644	1.2531	0.0042	1.0725

Table 5: Association rules without F-clustering (mapping table)

#	antecedents	count	consequents	count
1	Hypertensive heart disease	33624	Nonorganic sleep disorders	10956
2	Gonarthrosis [arthrosis of the knee]	9471	Recurrent depressive disorder	59941
3	Gastro-oesophageal reflux disease	7350	Recurrent depressive disorder	59941
4	Sleep disorders	14677	Recurrent depressive disorder	59941
5	Abdominal and pelvic pain	6755	Other anxiety disorders	42990
6	Gastro-oesophageal reflux disease	7350	Other anxiety disorders	42990
7	Sleep disorders	14677	Other anxiety disorders	42990
8	Cystitis	6711	Other anxiety disorders	42990
9	Sleep disorders	14677	Depressive episode	53034

4.3 Deep Learning

This subsection summarizes the results of publication [VII], which evaluates different deep learning approaches as DDSS decision technology. Deep learning is a subsymbolic AI method and is classified as a subset of ML. Deep learning uses several layers of artificial neural networks to mimic the way the human brain learns. We investigated the accuracy of non-sequential models like logistic regression (LR) and feed forward neural networks (FNN), sequential models like long short-term memory (LSTM) [45] and convolutional neural network [63] combined with LSTM (CNN-LSTM), and a gated recurrent unit [21] with a decay factor (GRU-decay), for predicting a psychiatric disorder. These were compared against our own novel Att-GRU-decay deep learning model, which additionally uses a self-attention layer [80] to detect hidden patterns in the patient's medical history that could indicate depression. Additionally, the decay factor helps to model the irregular times between events. We compare the average Area under the ROC Curve (AUC) and Area under the Precision-Recall Curve (AUPRC) scores obtained over 5-fold cross-validation. All results are reported in Table 6. Table 7 presents the specificity and sensitivity scores of the evaluated models. The attention layer of our proposed model allows visualization of disease patterns which were relevant to the depression prediction. This explainability component provides insight into how the network learns and why a certain output is generated. Fig. 4 shows an example plot of the attention filter. Important to note is that the plot only shows disease patterns learned by the attention layer and does not provide complete explainability for the GRU-decay part of the model. The prediction accuracy of deep learning not only outperforms all other approaches tested during this research, but our proposed Att-GRU-decay model also outperformed the current state of the art. The reasons for this include the power of feature generation of deep learning algorithms, which

Table 6: AUC and AUPRC scores on depression detection task over 5-fold cross validation ($\pm v$ denotes the standard deviation)

Models	AUC	AUPRC
LR	0.813 ± 0.002	0.296 ± 0.003
CNN-LSTM	0.849 ± 0.002	0.394 ± 0.009
LSTM	0.848 ± 0.001	0.385 ± 0.005
FNN	0.837 ± 0.002	0.374 ± 0.006
GRU-decay	0.989 ± 0.001	0.972 ± 0.001
Att-GRU-decay	$\mathbf{0.990 \pm 0.001}$	$\mathbf{0.974 \pm 0.002}$

Table 7: Specificity and sensitivity scores on the depression detection task

	Specificity		Sensitivity	
	Threshold			
	0.5	0.8	0.5	0.8
Models				
LR	0.705	0.960	0.787	0.263
CNN-LSTM	0.718	0.902	0.818	0.549
LSTM	0.724	0.916	0.814	0.523
FNN	0.714	0.911	0.810	0.528
GRU-decay	**0.995**	**0.999**	0.939	0.926
Att-GRU-decay	0.985	**0.999**	**0.955**	**0.944**

limits the amount of human intervention in data pre-processing and feature engineering, and the capabilities of more effective learning from complex data. Additionally, the best performing deep learning method in our evaluation took into consideration not only disease patterns but also the elapsed time between successive diagnoses. Additionally, the used dataset, which contained 26,973,943 diagnoses of 812,853 persons, strengthens the research results. One downside is that even though the attention filter allows for explainability of learned disease patterns relevant to the model output to a certain extent, full transparency as with a rule-based approach is not given. It is challenging to explain the contribution of each artificial neuron in the deep learning network and its individual importance in the downstream task. Additionally, massive data volumes for training are required since deep learning systems learn gradually. This training process is complex and demands a lot of computational power. A detailed analysis of the challenges in applying deep learning can be found at [66].

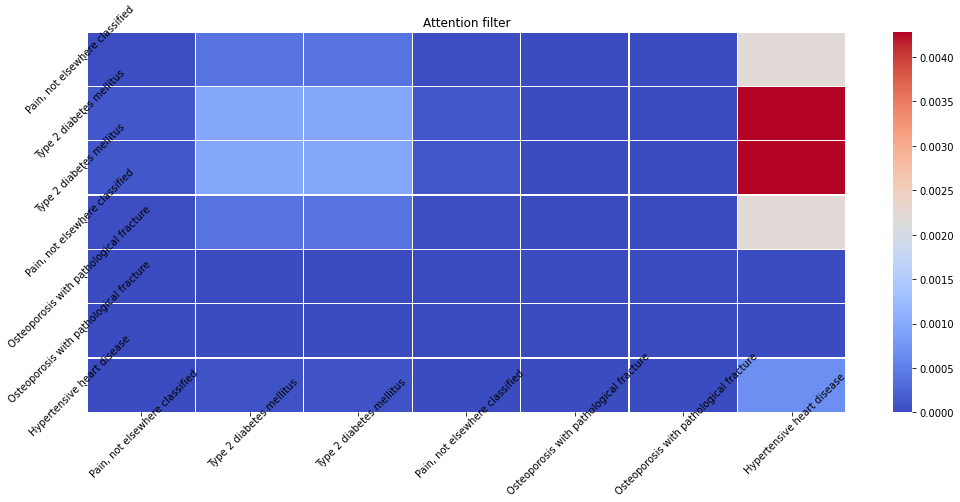


Figure 4: Example Attention Filter

5 Discussion

5.1 Outline of Research Findings and Evaluation

DDSSs in psychiatry have low adoption rates, and research generally has low maturity. This has been found in both publication [I], for DDSSs for PTSD, as well as in publication [II], for DDSSs for psychiatry as a whole. The average maturity level of research is 2.6, so mostly based on pure algorithm development. Additionally, the development and evaluation of these algorithms has mostly been done on small datasets (median sample size of 151.5). This raises questions about the trustworthiness of the claimed high accuracy rates of algorithms. Additionally, the clinical component, like the availability of data in the clinical workflow, the usefulness of the DDSS for clinicians, or user acceptance, has been widely neglected. All this influences low adoption rates. To overcome this, we propose a systematic approach for AI support in psychiatry that provides guidance on the IT perspective in addition to the clinical perspective. The framework created is based on (i) literature (publications [I] and [II]), (ii) expert knowledge obtained through a focus group interview (publication [III]), and (iii) practical experience from designing DDSSs based on real-world data (publications [IV], [V], [VI], and [VII]). This triangulation also allowed for a more thorough evaluation. While the focus group interview with a diverse group of nine DDSS experts evaluated the framework against practical usefulness, the scenario-based evaluation showed how the framework can actually be applied to raise DDSS adoption rates.

Data quality for AI-based DDSSs was evaluated in publication [IV] to get a better understanding of which features from what data source can be used for DDSS development. To find out which AI approaches are best suited to DDSS implementation, traditional machine learning-based [V], rule-based [VI] and deep learning-based [VII] approaches were investigated. These decision technologies were evaluated based on a real-world dataset consisting of longitudinal diagnostic data of nearly the whole Estonian population, including those suffering from a psychiatric disorder as well as a random, healthy sample. The deep learning approach based on our novel AttGRU-decay model outperformed not only the other approaches with an AUPRC of 0.974 but also the current state of the art. Additionally, our proposed rule-based approach showed promising results due to its flexibility,

its high explainability, and the transparency of the output.

5.2 Summary of Related Work

Many frameworks in the area of digital decision support have been published. One of the most cited domain agnostic frameworks for DDSS development is [75]. While they give a comprehensive general overview of the components of DDSSs, specifics from the healthcare domain are omitted to maintain a broader scope. One example of a domain-specific framework for the field of health care is the nonadoption, abandonment, scale-up, spread and sustainability (NASSS) framework [34], which has been cited frequently. However, the NASSS does not specifically deal with the challenges of AI in health care or addresses the practical shortcomings of DDSS development so far. As stated by Greenes et al., covering all aspects of decision support in health care in sufficient detail in a single overarching model is challenging [33]. Additionally, theoretical models that can be utilized for a wide range of scenarios often lose their usefulness for applied work. Therefore, there is a need for a multitude of frameworks for various aspects of the complex domain of decision support in health care. As for the DDSS prototype design, a compelling overview of related work on implementing DDSSs in psychiatry and a description of the existing shortcomings (like low maturity, low sample sizes, and a focus on pure algorithmic development under 'textbook conditions') can be found in publications [II] and [III].

5.3 Summary of Contribution

Gregor and Hevner propose that design science research, as used in this work, can result in two broad areas of contribution, namely design artifacts and design theories [35]. Both should provide not only an acceptable solution to a real-world problem but also novel inputs to knowledge [35]. In Sect. 3, we present the introduced framework for systematic AI support as design theory for designing useful DDSSs that create value in everyday clinical practice. This design theory has since been applied and tested by creating the DDSS prototypes described in Sect. 4. The prototypes can be seen as the design artifacts of this research.

From a more practical perspective, this research contributes to three main areas which were addressed by our research questions:

1. *Pointing out the shortcomings of current DDSS research*, namely the lack of interdisciplinarity for DDSS design and development, resulting in a sole focus on the algorithmic part of DDSSs, unreliable accuracy metrics because of low sample sizes for AI training and testing, and low maturity resulting in low clinical value of DDSSs.

2. *Proposing and assessing systematic AI support* based on our framework as a solution to raise DDSS adoption rates and increase their clinical benefit.

3. *Evaluating the current decision technologies* that can power DDSSs based on large amounts of real-world data; proposing a novel deep learning model that outperforms the current state of the art.

5.4 Limitations and Implications for Further Research

We do not propose that the systematic AI approach solves all problems concerning low adoption rates and low benefits of current DDSSs, but we see it as a step in the right direction in order to at least increase the maturity of prototypes from a technology point of view. Due to the technology focus of the proposed framework, dimensions around financial issues, marketing, and policy and political questions were omitted. These areas also have a high impact on DDSS success; therefore, further research is encouraged.

Furthermore, the generalizability of the proposed systematic AI approach and DDSS framework is suggested as an area of future research. Additional research on the validation of the proposed framework is encouraged as well. Another scenario-based evaluation using a successfully implemented DDSS like Duodecim's Evidence-Based Medicine electronic Decision Support (EBMEDS)[10] in Estonia is in planning. In order to assess whether the proposed DDSS scenario with the described decision technologies has a positive impact on patient outcome and/or the overall clinical process, a randomized control trial is needed. Applying the proposed decision technologies would raise the maturity level of the described DDSS scenario from 2 to 4.

Additionally, we only did an exemplary evaluation of promising algorithms for AI-based decision technologies. Since our Att-GRU-decay model performed so well, detailed research about other decision technologies like logic-based expert systems or hybrid approaches that combine symbolic and subsymbolic AI methods was considered out of scope. Depending on the randomized control trial results, the other decision technologies will be researched further. Especially the fusion of deep learning's ability to automatically extract intricate patterns from complex healthcare data with rule-based systems' capacity to incorporate domain-specific knowledge and enforce clinical guidelines holds high potential for enhancing DDSSs in health care.

One other limitation of this work, which also holds potential for further research, is that we only prototyped the data and decision technology dimension. Now, research on other domains of the framework is planned as part of TalTech eMed Lab's strategy. The data dimension offers the potential for further research. Currently, EHR systems mostly focus on recording disease data (as the data used in this research). However, recorded diseases are only an aggregation of symptoms. The aggregation of symptoms to one or more diseases can be wrong or biased. Especially with the high number of undiagnosed or misdiagnosed patients in psychiatry, the value of those data sources is questionable. The recording of symptoms itself seems to be a potential way to overcome the data quality challenge. Unobtrusive data collection (e.g. from wearables) offers a less biased way to gain insight into a patient's health and mental status.

6 Conclusion

Psychiatric disorders have a large impact on patients as well as society at large. Nevertheless, diagnostic accuracy remains low, leading to many people not receiving treatment for years. The success of AI could not yet be replicated in the psychiatric domain. There is a growing number of research articles proposing AI-based decision support, but maturity is generally low and research mostly focuses on the algorithmic part of DDSSs. Additionally, those AI algorithms are largely only trained and evaluated on small datasets. Low sample sizes make their claimed high accuracy rates questionable in a real-life usage scenario. This research contributes by (i) proposing a framework to raise DDSS adoption rates based on a systematic AI approach that takes into consideration the complete business process in which a DDSS should be applied, (ii) evaluating common decision technologies based on a large, real-world dataset to identify whether current AI algorithms are performing well enough, and (iii) proposing a novel deep learning algorithm that outperforms the current state of the art.

[10] https://www.ebmeds.org/en/

List of Figures

List of Tables

References

[1] A. Aboraya, E. Rankin, C. France, A. El-Missiry, and C. John. The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont)*, 3(1):41, 2006.

[2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[3] M. AlSalem, M. A. AlHarbi, A. Badeghiesh, and L. Tourian. Accuracy of initial psychiatric diagnoses given by nonpsychiatric physicians: A retrospective chart review. *Medicine*, 99(51), 2020.

[4] American Psychology Association. Depression assessment instruments, 2023. `https://www.apa.org/depression-guideline/assessment`, Last accessed on 2023-04-04.

[5] R. Anyoha. The history of artificial intelligence. *Science in the News*, 28, 2017.

[6] A. P. Association. and A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association Arlington, VA, 5th ed. edition, 2013.

[7] N. Bakkar, T. Kovalik, I. Lorenzini, S. Spangler, A. Lacoste, K. Sponaugle, P. Ferrante, E. Argentinis, R. Sattler, and R. Bowser. Artificial intelligence in neurodegenerative disease research: use of ibm watson to identify additional rna-binding proteins altered in amyotrophic lateral sclerosis. *Acta neuropathologica*, 135:227–247, 2018.

[8] J. Balch, G. R. Upchurch, A. Bihorac, and T. J. Loftus. Bridging the artificial intelligence valley of death in surgical decision-making. *Surgery*, 169(4):746–748, 2021.

[9] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571, 1961.

[10] M. Bertl. News Analysis for the Detection of Cyber Security Issues in Digital Healthcare: A Text Mining Approach to Uncover Actors, Attack Methods and Technologies for Cyber Defense. *Young Information Scientist*, 4:1–15, 2019.

[11] M. Bertl, N. Bignoumba, P. Ross, S. B. Yahia, and D. Draheim. Evaluation of Deep Learning-based Depression Detection using Medical Claims Data. *SSRN*, 2023.

[12] M. Bertl, K. J. I. Kankainen, G. Piho, D. Draheim, and P. Ross. Evaluation of Data Quality in the Estonia National Health Information System for Digital Decision Support. In *Proceedings of the 3rd International Health Data Workshop*. CEUR-WS, 2023.

[13] M. Bertl, T. Klementi, G. Piho, P. Ross, and D. Draheim. How Domain Engineering Can Help to Raise Adoption Rates of Artificial Intelligence in Healthcare. In *Proceedings of the 25th International Conference on Information Integration and Web-based Applications & Services*. Springer Nature, 2023.

[14] M. Bertl, J. Metsallik, and P. Ross. A Systematic Literature Review of AI-based Digital Decision Support Systems for post-traumatic Stress Disorder. *Frontiers in Psychiatry*, 13, 2022.

[15] M. Bertl, P. Ross, and D. Draheim. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *Database and Expert Systems Applications*, pages 104–111. Springer International Publishing, 2021.

[16] M. Bertl, P. Ross, and D. Draheim. A Survey on AI and Decision Support Systems in Psychiatry – Uncovering a Dilemma. *Expert Systems with Applications*, 202:117464, 2022.

[17] M. Bertl, P. Ross, and D. Draheim. Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors. *Health Policy and Technology*, 2023.

[18] M. Bertl, M. Shahin, P. Ross, and D. Draheim. Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 826–833. Association for Computing Machinery, 2023.

[19] K. Chockley and E. Emanuel. The end of radiology? three threats to the future practice of radiology. *Journal of the American College of Radiology*, 13(12):1415–1420, 2016.

[20] M. Chui, B. Hall, H. Mayhew, A. Singla, and A. Sukharevsky. The state of ai in 2022 — and a half decade in review, 2022. `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review`, Last accessed on 2023-04-04.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[22] V. Clarke, V. Braun, and N. Hayfield. Thematic analysis. *Qualitative psychology: A practical guide to research methods*, 3:222–248, 2015.

[23] R. M. Dawes, D. Faust, and P. E. Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

[24] N. K. Denzin. *The research act: A theoretical introduction to sociological methods*. Transaction publishers, 2017.

[25] J. Drescher. Out of dsm: Depathologizing homosexuality. *Behavioral sciences*, 5(4):565–575, 2015.

[26] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[27] Estonian National Institute for Health Development. Ra02: Residents with health insurance and health insurance coverage by sex and county - tervisestatistika ja terviseuuringute andmebaas, 2020. `https://statistika.tai.ee/pxweb/en/Andmebaas/Andmebaas__04THressursid__12Ravikindlustatud/RA02.px/`, Last accessed on 2022-03-04.

[28] E. Fast and E. Horvitz. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[29] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller. Watson: beyond jeopardy! *Artificial Intelligence*, 199:93–105, 2013.

[30] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter. Auto-Sklearn 2.0: The Next Generation. *arXiv:2007.04074 [cs, stat]*, 2020.

[31] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and Robust Automated Machine Learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Proc. of NIPS'2015 – the 28th Annual Conference on Neural Information Processing Systems*, pages 1–9, 2015.

[32] P. Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.

[33] R. A. Greenes, D. W. Bates, K. Kawamoto, B. Middleton, J. Osheroff, and Y. 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.

[34] T. Greenhalgh, J. Wherton, C. Papoutsi, J. Lynch, G. Hughes, S. Hinder, N. Fahy, R. Procter, S. Shaw, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*, 19(11):e8775, 2017.

[35] S. Gregor and A. R. Hevner. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, pages 337–355, 2013.

[36] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, et al. Cost of disorders of the brain in europe 2010. *European neuropsychopharmacology*, 21(10):718–779, 2011.

[37] M. Haenlein and A. Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4):5–14, 2019.

[38] A. Hamidia, F. Kheirkhah, M. Chehrazi, Z. Basirat, R. Ghadimi, S. Barat, P. Cuijpers, E. O'Connor, S. M. Mirtabar, and M. Faramarzi. Screening of psychiatric disorders in women with high-risk pregnancy: Accuracy of three psychological tools. *Health Science Reports*, 5(2):e518, 2022.

[39] M. Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1):56, 1960.

[40] E. J. Hegedus and J. Moody. Clinimetrics corner: the many faces of selection bias. *Journal of Manual & Manipulative Therapy*, 18(2):69–73, 2010.

[41] M. Henderson, S. B. Harvey, S. Øverland, A. Mykletun, and M. Hotopf. Work and common psychiatric disorders. *Journal of the Royal Society of Medicine*, 104(5):198–207, 2011.

[42] A. Hevner and S. Chatterjee. Design science research in information systems. In *Design research in information systems*, pages 9–22. Springer, 2010.

[43] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[44] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1):58–64, 2000.

[45] S. Hochreiter, J. Schmidhuber, et al. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[46] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[47] M. Indulska and J. Recker. Design science in IS research : a literature analysis. In *Information systems foundations: The role of design science*, pages 285–302. ANU Press, 2008.

[48] V. Kaul, S. Enslin, and S. A. Gross. History of artificial intelligence in medicine. *Gastrointestinal endoscopy*, 92(4):807–812, 2020.

[49] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. 2007.

[50] K. Kõnd and A. Lilleväli. E-prescription success in estonia: The journey from paper to phamacogenomics. *Eurohealth*, 25(2):18–20, 2019.

[51] J. H. Krystal and M. W. State. Psychiatric disorders: diagnosis to therapy. *Cell*, 157(1):201–214, 2014.

[52] A. E. Lake III, J. C. Rains, D. B. Penzien, and G. L. Lipchik. Headache and psychiatric comorbidity: historical context, clinical implications, and research relevance. *Headache: The Journal of Head and Face Pain*, 45(5):493–506, 2005.

[53] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1–e34, 2009.

[54] K. Mak, H. C. Pilles, M. Bertl, and J. Klerx. Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie. *Schriftenreihe der Landesverteidigungsakademie. Wien: BM für Landesverteidigung und Sport*, 2018.

[55] N. McCauley and M. Ala. The use of expert systems in the healthcare industry. *Information & Management*, 22(4):227–235, 1992.

[56] E. A. McGlynn, S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, and E. A. Kerr. The quality of health care delivered to adults in the united states. *New England journal of medicine*, 348(26):2635–2645, 2003.

[57] C. McIntosh, editor. *Cambridge Learner's Dictionary*. Cambridge University Press, 2012.

[58] P. E. Meehl. When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, 4(4):268, 1957.

[59] Y. Mintz and R. Brodie. Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, 28(2):73–81, 2019.

[60] A. J. Mitchell, A. Vaze, and S. Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619, 2009.

[61] S. A. Montgomery and M. Åsberg. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, 134(4):382–389, 1979.

[62] P. L. Morselli and R. Elgie. Gamian-europe*/beam survey i–global analysis of a patient questionnaire circulated to 3450 members of 12 european advocacy groups operating in the field of mood disorders. *Bipolar Disorders*, 5(4):265–278, 2003.

[63] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[64] F. Rabiee. Focus-group interview and data analysis. *Proceedings of the nutrition society*, 63(4):655–660, 2004.

[65] D. Rhodes, K. F. McFarland, W. H. Finch, and A. O. Johnson. Speaking and interruptions during primary care office visits. *Family medicine*, 33(7):528–532, 2001.

[66] N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, et al. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, 2022.

[67] V. Sauter. *Decision support systems: an applied managerial approach*. John Wiley & Sons, Inc., 1997.

[68] W. B. Schwartz. Medicine and the computer: the promise and problems of change. *Use and impact of computers in clinical medicine*, pages 321–335, 1987.

[69] R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's Paradox: A Step Towards Fairness in Machine Learning. In *European Conference on Advances in Databases and Information Systems*, pages 67–76. Springer, 2022.

[70] M. K. Shaw, S. A. Davis, A. B. Fleischer Jr, and S. R. Feldman. The duration of office visits in the united states, 1993 to 2010. *The American Journal of Managed Care*, 2014.

[71] A. M. Shin, I. H. Lee, G. H. Lee, H. J. Park, H. S. Park, K. I. Yoon, J. J. Lee, and Y. N. Kim. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthcare informatics research*, 16(2):77–81, 2010.

[72] J. Simpson, editor. *Oxford English Dictionary*. Oxford University Press, 2000.

[73] J. O. Sines. Actuarial versus clinical prediction in psychopathology. *The British Journal of Psychiatry*, 116(531):129–144, 1970.

[74] T. Singh and M. Rajput. Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)*, 3(10):57, 2006.

[75] R. H. Sprague Jr. A framework for the development of decision support systems. *MIS quarterly*, pages 1–26, 1980.

[76] D. J. Stein, K. A. Phillips, D. Bolton, K. Fulford, J. Z. Sadler, and K. S. Kendler. What is a mental/psychiatric disorder? from dsm-iv to dsm-v. *Psychological medicine*, 40(11):1759–1765, 2010.

[77] E. Strickland. Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE Spectrum*, 56(4):24–31, 2019.

[78] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

[79] A. Thamba and R. B. Gunderman. For watson, solving cancer wasn't so elementary: prospects for artificial intelligence in radiology. *Academic Radiology*, 29(2):312–314, 2022.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[81] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, et al. Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2163–2196, 2012.

[82] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

[83] S. Z. Williams, G. S. Chung, and P. A. Muennig. Undiagnosed depression: A community diagnosis. *SSM-Population Health*, 3:633–638, 2017.

[84] H.-U. Wittchen, F. Jacobi, J. Rehm, A. Gustavsson, M. Svensson, B. Jönsson, J. Olesen, C. Allgulander, J. Alonso, C. Faravelli, et al. The size and burden of mental disorders and other disorders of the brain in europe 2010. *European neuropsychopharmacology*, 21(9):655–679, 2011.

Acknowledgements

Completing this academic journey has been a challenging yet fulfilling endeavor, and I owe a debt of gratitude to many individuals who have supported and inspired me along the way. First and foremost, I want to express my deepest appreciation to my grandmother, Elfriede Hackl, who passed away this year. Her unconditional love, unwavering support, and words of wisdom have been a constant source of strength throughout this academic pursuit. Grandmother, your belief in me and your encouragement during both triumphs and setbacks have been invaluable. I am also immensely grateful to my family for their endless support. To my parents, thank you for your understanding, encouragement, and for creating a home environment that fostered my academic ambitions. Amid the challenges of research, the camaraderie and laughter of my friends were indispensable. Their friendship not only provided moments of respite but also served as a vital source of balance in my life. I extend my gratitude to my colleagues and peers, whose insights and discussions have enriched my research. The collaborative spirit within our academic community has undoubtedly contributed to the depth and breadth of this work. Special thanks are due to the Samaritan Austria Rapid Response Team, of which I am privileged to be a part. Your commitment to healthcare and emergency response not only provided me with invaluable insights into the intricacies of the field but also played a crucial role in shaping both my research direction and my personal growth. The experiences and lessons learned during our collective efforts have been transformative, and I am grateful for the opportunity to contribute to such a dedicated and compassionate team. I am indebted to my supervisors, Prof. Peeter Ross and Prof. Dirk Draheim, for their guidance, expertise, and unwavering support. Their mentorship has been instrumental in shaping the direction of my research and in helping me navigate the complexities of the academic journey. To everyone who has played a role, big or small, in my academic and personal growth, thank you. This thesis represents not only my individual effort but also the collective support and encouragement of a wonderful network of people. I am truly fortunate to have such incredible individuals in my life. Finally, I dedicate this thesis to my family, friends, colleagues, and especially to my grandmother, whose love and support have been the foundation upon which this academic achievement stands.

Abstract
Systematic AI Support for Psychiatry: A Framework on How to Implement Decision Support Systems

Diseases of the brain are, with an annual prevalence of 38%, not only very common, but they also account for more than €461 billion in healthcare costs in Europe. These diseases are often diagnosed late or not at all. Artificial Intelligence (AI) could improve the diagnostic process. Two systematic literature reviews containing information from 80 research papers on Digital Decision Support Systems (DDSSs) in psychiatry show low adoption rates and generally low maturity. Research mostly focuses on pure algorithm development, and evaluation is performed on small datasets. This raises questions about the trustworthiness of the claimed high accuracy rates of algorithms. Additionally, the clinical component, such as the availability of data in the clinical workflow, the usefulness of the DDSS for clinicians or user acceptance, has been widely neglected. All this influences low adoption rates. This research proposes a systematic approach that takes into consideration both the clinical and the technical aspects. For this systematic AI support, a framework with dimension data, technology, user group, medical domain, decision, validation and maturity serves as a tool for more holistic DDSS development. The framework was derived based on (i) literature, (ii) data from a focus group interview with nine DDSS experts from various fields, and (iii) practical experience. A scenario-based evaluation and a focus group interview were used to evaluate the framework. To overcome the potential issue of insufficiently working decision technology, traditional machine learning algorithms, a rule-based approach and several deep learning methods, including our own novel attGRU-decay model, were benchmarked on real-world diagnostic data from 812,853 patients with a total of 26,973,943 diagnoses. Our attGRU-decay model outperformed the other methods and the current state of the art with an AUPRC of 0.974.

These results can be clustered into three main contributions:

1. *Pointing out the shortcomings of current DDSS research*, namely the lack of interdisciplinarity for DDSS design and development, resulting in a sole focus on the algorithmic part of DDSSs, unreliable accuracy metrics due to low sample sizes for AI training and testing and low maturity and low clinical value of DDSSs.

2. *Proposing and assessing systematic AI support* based on our framework as a solution to raise DDSS adoption rates and increase their clinical benefit.

3. *Evaluating the current decision technologies* that can power DDSSs based on large amounts of real-world data; proposing a novel deep learning model that outperforms the current state of the art.

Kokkuvõte
Tehisintellekti süstemaatiline kasutamine psühhiaatrias: otsustustoe rakendamist toetav raamistik

Ajutegevusega seotud haigused, mille aastane levimus on 38%, ei ole mitte ainult väga levinud haigused, vaid tekitavad Euroopas rohkem kui 461 miljardi euro väärtuses tervishoiukulusid. Nende haiguste diagnoosini jõutakse sageli hilja või need jäävad üldse diagnoosimata. Tehisintellekt (TI) võib aidata seda olukorda parandada. Kaks süstemaatilist kirjanduse ülevaadet, mis sisaldavad teavet digitaalsete otsustustugede (inglise keeles – Digital Decision Support Systems (DDSS)) kohta 80 uurimistööst, näitavad nende vähest kasutuselevõttu ja üldiselt madalat küpsustaset. Käesoleval ajal tehtav teadustöö keskendub enamasti ainult algoritmi arendamisele ja otsustustoe hindamine toimub väikeste andmekogumite põhjal. See tekitab küsimusi algoritmide väidetava suure täpsuse usaldusväärsuse kohta. Lisaks on kliiniline komponent nagu näiteks andmete kättesaadavus kliinilises töövoos, DDSS-i kasulikkus arstide jaoks või kasutajate poolne aktsepteerimine laialdaselt uurimistöös tähelepanuta jäetud. Kõik see on põhjuseks, miks DDSS-i kasutuselevõtt on vähene. Käesolev uurimustöö pakub välja süstemaatilise lähenemise, mis võtab DDSS-i arendamisel arvesse nii meditsiinilisi kui ka tehnoloogilisi aspekte. Selleks, et süsteemselt toetada tehisintellekti rakendamist terviklikuma DDSS-i arendamiseks töötati välja raamistik, mis vaatleb eraldi komponentidena andmeid, tehnoloogiat, kasutajarühmi, meditsiinivaldkonda, otsuseid, valideerimist ja küpsusastet. Raamistik tuletati (i) kirjanduse, (ii) üheksa erineva valdkonna DDSS-eksperdi fookusgrupi intervjuu andmetel ja (iii) praktilisel kogemusel. Raamistiku hindamiseks kasutati stsenaariumipõhist hindamist ja fookusgrupi intervjuud. Et lahendada väidetavalt meditsiinis seni ebapiisavalt rakendatud otsustustoe probleemi võrreldi traditsioonilisi masinõppe algoritme, reeglipõhist lähenemisviisi ja mitmeid süvaõppe meetodeid, sealhulgas meie enda uudset attGRU-decay meetodit kasutades 812 853 patsiendi, kellel oli kokku 26 973 943 diagnoosi, tegelikke diagnostilisi andmeid. Meie attGRU-decay ületas teisi meetodeid ja praegust parima praktika taset AUPRC-ga 0,974.

Doktoritööl on kolm peamist tulemust:

1. Uuring toob välja, et käesoleval ajal DDSS-i kohta tehtava teadustöö puuduseks on selle kavandamise ja arendamise interdistsiplinaarsuse puudumine, mille põhjuseks on keskendumine ainult DDSS-i algoritmidele. Samuti ebausaldusväärsed täpsusmõõdikud väikese valimi tõttu TI koolitamisel ja testimisel ning DDSS-ide madal küpsusaste ja kliiniline väärtus.

2. Suurendamaks DDSS-i kasutuselevõttu ja kliinilist kasu pakutakse välja raamistik, mis toetab tehisintellekti süsteemset kasutamist ja hindamist.

3. Hinnatakse praeguseid otsustustoes kasutatavaid TI meetodeid, mis suurandmeid kasutades võivad suurendada DDSS-i kasu, ja pakutakse välja käesoleval ajal kasutatavaid meetodeid ületav uudne süvaõppe mudel..

Appendix 1

[I]

M. Bertl, J. Metsallik, and P. Ross. A Systematic Literature Review of AI-based Digital Decision Support Systems for post-traumatic Stress Disorder. *Frontiers in Psychiatry*, 13, 2022

frontiers | Frontiers in Psychiatry

TYPE Systematic Review
PUBLISHED 09 August 2022
DOI 10.3389/fpsyt.2022.923613

OPEN ACCESS

EDITED BY
Oswald David Kothgassner,
Medical University of Vienna, Austria

REVIEWED BY
Weihui Li,
Central South University, China
Karin Waldherr,
Ferdinand Porsche FernFH – Distance Learning University of Applied Sciences, Austria

*CORRESPONDENCE
Markus Bertl
mbertl@taltech.ee

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Psychiatry

RECEIVED 19 April 2022
ACCEPTED 15 July 2022
PUBLISHED 09 August 2022

CITATION
Bertl M, Metsallik J and Ross P (2022) A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder.
Front. Psychiatry 13:923613.
doi: 10.3389/fpsyt.2022.923613

A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder

Markus Bertl*, Janek Metsallik and Peeter Ross

Department of Health Technologies, School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia

Objective: Over the last decade, an increase in research on medical decision support systems has been observed. However, compared to other disciplines, decision support systems in mental health are still in the minority, especially for rare diseases like post-traumatic stress disorder (PTSD). We aim to provide a comprehensive analysis of state-of-the-art digital decision support systems (DDSSs) for PTSD.

Methods: Based on our systematic literature review of DDSSs for PTSD, we created an analytical framework using thematic analysis for feature extraction and quantitative analysis for the literature. Based on this framework, we extracted information around the medical domain of DDSSs, the data used, the technology used for data collection, user interaction, decision-making, user groups, validation, decision type and maturity level. Extracting data for all of these framework dimensions ensures consistency in our analysis and gives a holistic overview of DDSSs.

Results: Research on DDSSs for PTSD is rare and primarily deals with the algorithmic part of DDSSs (n = 17). Only one DDSS was found to be a usable product. From a data perspective, mostly checklists or questionnaires were used (n = 9). While the median sample size of 151 was rather low, the average accuracy was 82%. Validation, excluding algorithmic accuracy (like user acceptance), was mostly neglected, as was an analysis concerning possible user groups.

Conclusion: Based on a systematic literature review, we developed a framework covering all parts (medical domain, data used, technology used for data collection, user interaction, decision-making, user groups, validation, decision type and maturity level) of DDSSs. Our framework was then used to analyze DDSSs for post-traumatic stress disorder. We found that DDSSs are not ready-to-use products but are mostly algorithms based on secondary datasets. This shows that there is still a gap between technical possibilities and real-world clinical work.

KEYWORDS

decision support systems (DSS), post-traumatic stress disorder (PTSD), artificial intelligence (AI), machine learning (ML), systematic literature review (SLR), clinical decision support (CDS), psychiatry, mental health

Introduction

According to Sauter, Digital Decision Support Systems (DDSSs) are computer-based systems that bring together information from various sources, assist in the organization and analysis of information and facilitate the evaluation of assumptions underlying the use of specific models (1). The concept of decision support systems originated in the 1960s (2) when researchers began to study computerized methods to assist in decision-making (3–5). Since then, the idea has extended throughout a broad spectrum of domains, one of which is healthcare. This work focuses on decision support systems in mental health, more precisely on decision support systems for PTSD. The American Psychiatric Association defines PTSD as "a psychiatric disorder that can occur in people who have experienced or witnessed a traumatic event such as a natural disaster, a serious accident, a terrorist act, war/combat, rape or other violent personal assault" (6). People with PTSD experience recurrent thoughts about their traumatic experience that influence their daily life. The lifetime prevalence of PTSD is around 12.5% (7). However, people suffering from PTSD are often undiagnosed or misdiagnosed, resulting in incorrect, incomplete or missing treatment (8). To investigate whether DDSSs could be a solution to this problem, we aim to review available decision support systems for PTSD and map their technological approaches in order to understand possible research gaps and obstacles in introducing decision support systems to clinical processes. Since no available reference architecture for decision support systems is applicable to our research, we contribute by introducing a novel framework for decision support systems that can be used to analyze existing systems. Ultimately, this also accelerates the development of new systems by highlighting essential dimensions.

Designers of earlier DDSSs have applied multiple alternative approaches for converting real-world data into something that stimulates better decisions. Information-management-based DDSSs try to organize data into usable presentations; modeling-(or data-analytics)-based DDSSs attempt to apply statistical (learning) methods for finding patterns or calculating indicators; and knowledge-management-based systems apply externally prepared algorithms (expert rules) to find matching data or derive new facts (9). While AI has been an essential element of DDSSs throughout its history, only recently has a new generation of decision support been facilitated by the availability of powerful computing tools to properly manage big data and to analyze and generate new knowledge. The evaluation of AI's earlier implementations was limited to the design and development phase; machine learning-based algorithms often do not generalize beyond the training data set (10). However, studies have still shown the benefits of machine learning algorithms in DDSSs (11–13). Current studies that test the application of healthcare AI algorithms often omit details of DDSS tools that apply AI models. A well-designed DDSS is likely to enable the real-world application of AI technology (14).

This review aims to contribute by introducing a framework for the features of DDSS implementation in mental health. We aim to identify the prevalent features of the current state of research on DDSS. Often, the development of information systems involves the continuous introduction of new features and quality improvements. We hypothesized that each available article presents only a selection of features, a selection which is dependent on the maturity of the DDSS. Maturity models are increasingly used as a means of benchmarking or self-assessment of development (15). In healthcare informatics, many maturity models are available [e.g., Hospital Information System Maturity Model (16)], but none of these models strictly provides an informed approach for the assessment of research on decision support systems (17). The available maturity models instead tend to look at the level of organizational adoption of specific technologies (e.g., how much an organization values data analytics technology) and provide little support for deciding on the readiness of DDSS tools in their early phases of development. As AI is often an essential element of a DDSS, we also explored AI maturity models. AI maturity models mostly look into the level of AI adoption in an organization rather than the maturity of the AI technology itself (18–20).

A DDSS is not a single technology but rather a set of integrated technologies (21–25). Sauser et al. (26) suggested a measure of System Readiness Level (SRL), which expresses the level of maturity of a system consisting of a set of integrated technologies (26). Exploring AI technology readiness or maturity, we encountered suggestions to look separately into the AI system's capacities of integrating existing data sources (machine-machine intelligence), interacting with human users (human-computer intelligence) and applying intelligent reasoning (core cognitive intelligence) (27).

Methods

To have a transparent and objective approach for this literature review, we decided to apply the five stages suggested by Kitchenham's "Guidelines for performing Systematic Literature Reviews in Software Engineering" (28):

(1) Search Strategy
(2) Study Selection
(3) Study Quality Assessment
(4) Data Extraction
(5) Data Synthesis

Research questions

Since our aim is to understand current research on decision support systems for PTSD, this paper is based on two research questions. First, we look for state-of-the-art decision support systems for post-traumatic stress disorder (RQ1). Second, we investigate the component elements of current decision support systems for PTSD (RQ2).

Search strategy

We built a search string based on the research questions identified and applied it to the Scopus abstract and citation database. Scopus was chosen as the primary source because it is the largest abstract and citation database of research literature with 100% MEDLINE coverage (29). The initial search string consisted of the disease to investigate – post-traumatic stress disorder – its abbreviation PTSD as well as the term "decision support." To find papers that covered the prediction and classification of PTSD, we also added Artificial Intelligence. In Scopus, we applied the search string to the title, abstract and tags of the research papers. We restricted our search to only include journal articles or conference proceedings in English. We also conducted a manual search using Google Scholar and the web to find additional research; however, this did not bring up any new articles not already covered by our database search and our reference screening process. We formed our search criteria as ("decision support" OR "Artificial Intelligence") AND [PTSD OR (post AND traumatic AND stress AND disorder)].

We conducted the search in Scopus on 3 March 2021. It resulted in 75 papers; reference screening of the included literature brought up an additional 13 papers. Our search process is visualized in **Figure 1**.

Study selection

The titles and abstracts of the queried articles were analyzed to identify relevant articles from the results of the search string queries. Articles fitting the research questions and meeting the inclusion criteria (see section "inclusion criteria") as well as the quality criteria (see section "study quality assessment") were included. Since the goal of this research is to give an overview of the state of the art, we did not put any constraints on study types and designs. To reduce bias in the study selection process, the task was done by two researchers independently. The two result sets were then merged and deviations were discussed among the authors. This resulted in a total set of 17 research papers.

We then repeated this process step to extract relevant studies from the reference lists of the selected articles. This resulted in 13 new research papers.

Inclusion criteria

Table 1 presents the inclusion criteria applied to the articles in our review (Inclusion criteria).

Study quality assessment

Table 2 presents the inclusion criteria applied to the articles in our review (Quality criteria).

Data extraction and synthesis

Data extraction and synthesis were based on an inductive approach. We applied thematic analysis (30) to answer our research questions. First, clear, scoped questions for data extraction were formed. Two researchers read through all the articles and iteratively clustered all of the information available on decision support systems into the extraction parameters. These extraction parameters describe how decision support systems work. This process is shown in **Figure 2**.

The answers extracted from the EQs (see **Table 3**) were then combined upon the agreement of the authors to create a feature matrix. The extracted features were then further clustered to create a common terminology that allows further analysis and the possibility to compare results. In the end, we combined the developed extraction questions and the clustered scales of each question into a novel framework for decision support systems in mental health.

Results

The selected 30 research articles (31–60) were published between 2001 and 2019. Three articles were published in journals about medical informatics, 10 in computer science journals or proceedings and 17 in medical journals. The following table shows how often each extraction parameter was present and indicates the terminology used in the selected studies. The terminology shown in **Table 4** was developed by manual, iterative clustering of the extracted features until the authors were satisfied with the granularity.

A framework for digital decision support systems

Based on our aim to find all relevant features of decision support systems in the PTSD area and our systematic literature review results, we propose a multidimensional framework that covers the different areas of DDSS. Each dimension represents one of our extraction parameters. **Figure 3** illustrates our framework with the different dimensions of DDSSs. Based on the extracted data, we clustered the terminology

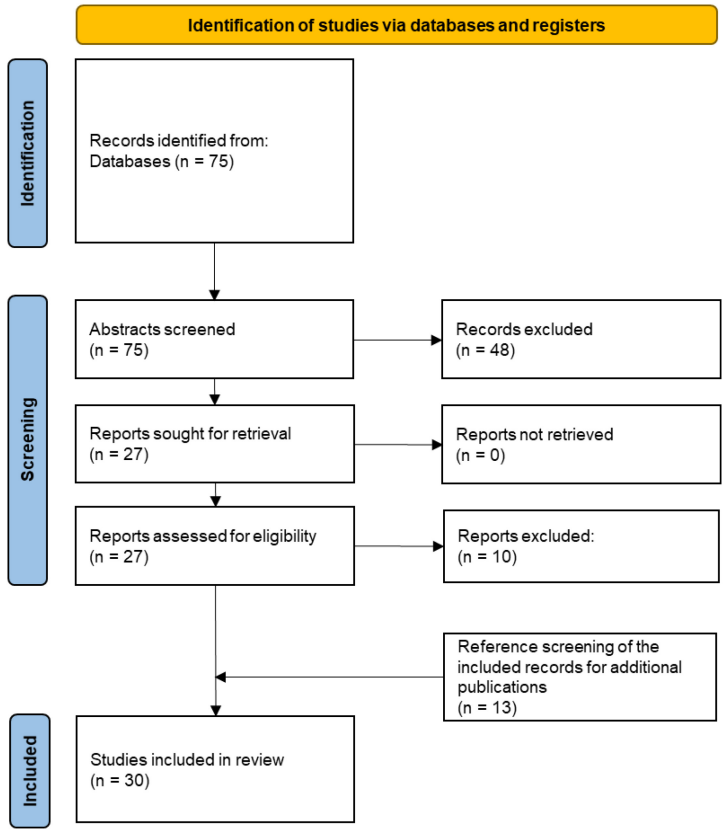


FIGURE 1
Search strategy.

to develop scales for dimensions in order to make results better analyzable.

Input Data: The input data dimension defines the information needed by a decision support system in order to function. Possible data could be structured like socio-demographic information or coded data [for example, with the International Statistical Classification of Diseases and Related Health Problems (ICD) (61) or the Diagnostic and Statistical Manual of Mental Disorders (DSM) (62)] as well as semi-structured information like patient records or unstructured information like free text or medical images. A combination of different structured, semi-structured and/or unstructured data is also possible.

TABLE 1 Inclusion criteria.

#	Inclusion criteria
IC1	Does the study deal with decision support systems (e.g., systems that help diagnose, screen, predict or treat)
IC2	Does this study apply a computerized algorithm?
IC3	Does this article deal with PTSD?
IC4	Is the article related to at least one of our research questions?

TABLE 2 Quality criteria.

#	Quality criteria
QC1	Is the research a journal article or conference proceeding?
QC2	Is the research peer-reviewed?
QC3	Does the study have a well-defined structure?
QC4	Does the study bring evidence for the proposed approach (either by citing relevant literature or validating the results)?
QC5	Does the study have ethics approval (if required by the study design)?

Technology: The technology dimension describes how the decision support system is implemented. This involves three sub-dimensions:

Decision technology: The decision technology explains the intelligence of the cognition of the system. This is the algorithm that powers the decision-making. Examples are different machine learning algorithms such as support vector machines or other statistical methods as well as rule-based approaches.

Interaction technology: This sub-dimension describes the technology needed to interact with other systems or user groups in the clinical process. Interaction technology can be API-based interfaces to systems, graphical user interfaces (websites, mobile apps) or sensory input like conversational interfaces (chatbots).

Data collection technology: The data collection technology sub-dimension defines how the data described in the input data dimension are collected. Examples are instance sensors, questionnaires or chatbots.

Validation: Validation describes how the success of decision support systems is measured.

Accuracy: The decision support system is evaluated by how many right or wrong decisions it makes. Examples are accuracy, recall (sensitivity), precision, specificity, area under the curve (AUC) values and F1 scores (harmonic mean of recall and precision).

User acceptance: End-users are involved in the evaluation of the DDSS.

Efficacy: The impact of the decision support system is evaluated based on potential benefits.

Security: The DDSS is evaluated against security regulations.

Legal: The legal compliance of the DDSS is evaluated.

User group: This dimension captures the different user groups interacting with the decision support system in the clinical process.

TABLE 3 Extraction questions (EQ).

#	Extraction parameters
EQ1	On the basis of which input data do existing decision support systems in mental health operate?
EQ1.2	What was the data sample size?
EQ2	What is the implementation technology of the DDSS?
EQ2.1	Decision technology
EQ2.2	User Interaction/Interface/Application
EQ2.3	Data collection technology
EQ3	What feature was validated?
EQ4	Which user groups are involved in the use of DDSS in mental health?
EQ5	What diseases are currently targeted by DDSS in mental health?
EQ6	What decisions are supported by the system?
EQ7	What maturity level does the DDSS have?

Medical domain: The medical domain dimension describes the disease for which the decision support system can be applied.

Decision: The following scale defines the decisions a digital decision support system can support:

Prediction: The system outputs a risk score based on the likelihood that someone gets a disease.

Assessment: The patient is already sick (knowingly or unknowingly).

Diagnosis: Testing individuals with symptoms and/or suspicion of illness

Screening: Testing for individuals without specific symptoms

Monitoring: Decision support that evaluates symptom severity or treatment progress

Treatment: Recommendation or intervention concerning care or therapy

Maturity: As none of the existing maturity models fits our research, we designed a DDSS maturity model based on the SLR scale (26), but with adaptions specific to healthcare. It introduces additional gradation for noticing the moment where human interaction is added to the core AI algorithm. Our maturity

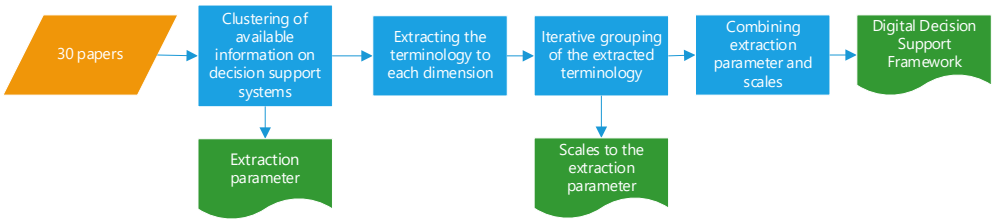


FIGURE 2
Extraction process.

levels describe on a scale from one to seven how advanced the DDSS is. Not all of the abovementioned dimensions are necessarily present in each of the maturity levels. As the maturity level gets higher, more dimensions are described.

1. Idea without implementation
2. Implementation without real-world interaction (algorithm development)
3. Implementation with real-world interaction but without patient intervention
4. Fully functioning prototype, system triggers real-world action, e.g., clinical trial
5. Operational product (at least one adopter, certified if required)
6. Locally adopted product
7. World-wide adopted product (transformational).

Data synthesis input data (EQ1)

The data used by digital decision support systems in the context of PTSD is diverse. Voice data (35, 45, 46, 55), text data (38, 48, 50), checklists and questionnaires (32, 33, 37, 41–43, 52, 53, 59), bio signals (32, 33, 36, 44, 45, 51, 57) and electronic medical records (34, 47, 56) as well as secondary data from other clinical studies (31, 40, 49, 54) are used. One article used the choices made by a virtual avatar in a role-playing game as input data (39). Of the 30 publications included in this review, 28 mentioned the sample size of the data they used to develop and test their decision support system. The minimum sample size was 10, and the maximum was 89,840 with a median (IQR) m = 151.5 (54.25 to 656.25). The violin plots (**Figures 4**, **5**) below show the distribution of the sample size. The top three outliers (89,840; 89,840; 5,972) were neglected in **Figure 5** for better visibility.

Figure 6 shows the data dimension of the studies in our review and indicates how the data used correlate with the average maturity levels of the DDSS. It visualizes the frequency and maturity of DDSSs based on the different data sources.

Data synthesis implementation (EQ2)

The majority (n = 15) of the investigated research uses a neural network approach (including support vector machines) in their systems. In 11 cases, support vector machines (SVM) were used. Other algorithms used were regressions, decision trees, random forest and rule-based approaches. We observed that 20 research papers did not have or mention any user interaction but worked solely on secondary data. The others used questionnaires or surveys, virtual humans or virtual reality. McWorther et al. proposed using temperature control, aromatherapy and auditory therapy capabilities for user interaction (36). Concerning maturity levels, AI algorithms are still mostly on maturity level two. Most advanced in terms of maturity were statistical methods and text mining methods, as indicated in **Figure 7**. The categories "statistics" and "machine learning" (ML) arose because some studies mentioned only these broad categories without further specifics.

Data synthesis validation (EQ3)

The majority (n = 23) of articles validated the accuracy of the DDSS studied. Three articles validated user acceptance, two validated efficacy and three did not mention validation. Comparing algorithmic validation among research papers was difficult since a variety of scores, such as F1 scores, area under the receiver operating curve (63) or overall accuracy, were used and they cannot be converted. To be able to provide an estimation of how well current DDSSs perform, we extracted all accuracy measurements present in each paper and aggregated each scale individually. The mean accuracy (n = 11) of the DDSSs is μ = 82.2% with a median of η = 82% and a standard deviation of σ = 0.095. The mean area under the curve value (n = 8) is μ = 0.845 with a median of η = 0.84 and a standard deviation of σ = 0.064.

Data synthesis user groups (EQ4)

The user groups mentioned were patients, clinicians and supporters of patients; however, the majority of papers did not explicitly mention specific user groups for their systems. Research covering decision support systems with higher maturity levels (four and above) included this information. Research dealing with decision support systems with lower maturity often lacked a clear user group since the process of using the proposed systems was not defined at that stage.

Data synthesis medical domain (EQ5)

In addition to PTSD, which was tackled by all 30 research papers, four investigated depression (46–48, 55), two anxiety (34, 48) and one paranoia (58).

Data synthesis decisions supported (EQ6)

Research focusing on predicting PTSD or its symptoms was most common (n = 11). Six papers focused on screening (35, 38, 45, 46, 50, 55) and six on treatment (32, 36, 43, 51, 53, 56). Four

TABLE 4 Terminology extraction.

EQ	Number of mentions	Terminology (frequency)
1 – Data	30	Jerusalem Trauma Outreach and Prevention Study (3); checklist (5); questionnaire (4); speech data (4); text data (3); electronic health records (3); sensor data (6); reactions in VR (2)
1.1 – Sample size	28	Not applicable (quantitative features)
2.1 – Decision technology	27	Machine learning algorithm; feed forward neural network; support vector machines, random forest; decision tree; sequential minimal optimization (SMO); Naïve Bayes; logistic regression; text mining; (LIWC); rule based
2.2 – Interaction technology	24	Questions (3); temperature control (1); aromatherapy (1); auditory therapy (1); virtual human (2); online survey (1); role-play-game (1); virtual reality (2)
2.3 – Data collection technology	22	Mobile app (4); web portal (3); skin conductance sensor (1); heart rate (1); accelerometer (1); IoT devices (1); microphone (1); webcam (1); Kinect (1); VR headset (1)
3 – Validation	29	Accuracy (23); user acceptance (3); efficacy (2)
4 – User groups	12	Patients (10); supporters (1); clinicians (6)
5 – Disease	30	PTSD (30); depression (4); anxiety (1); PTSD comorbidities (1); paranoia (1)
6 – Decisions	29	Prediction (11); assessment (1); diagnosis (4); screening (6); monitoring (5); treatment (6)
7 – Maturity level	30	Not applicable (quantitative features)

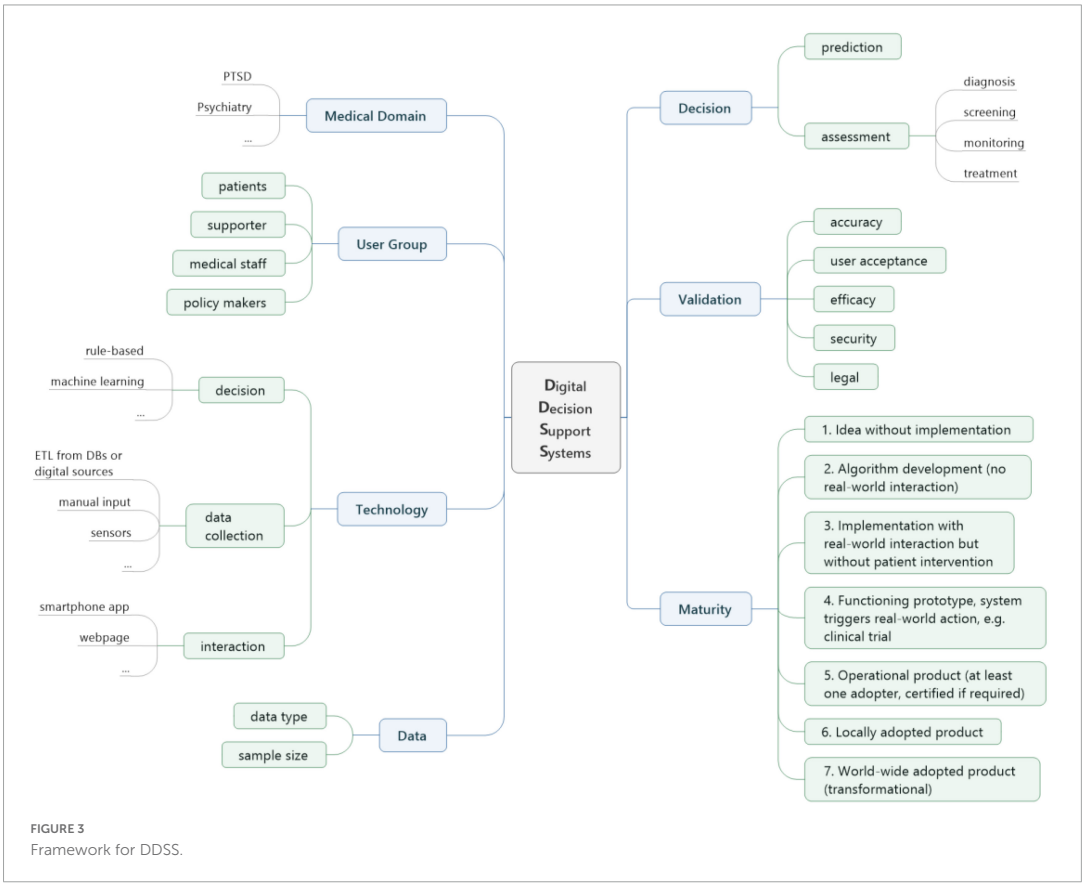


FIGURE 3
Framework for DDSS.

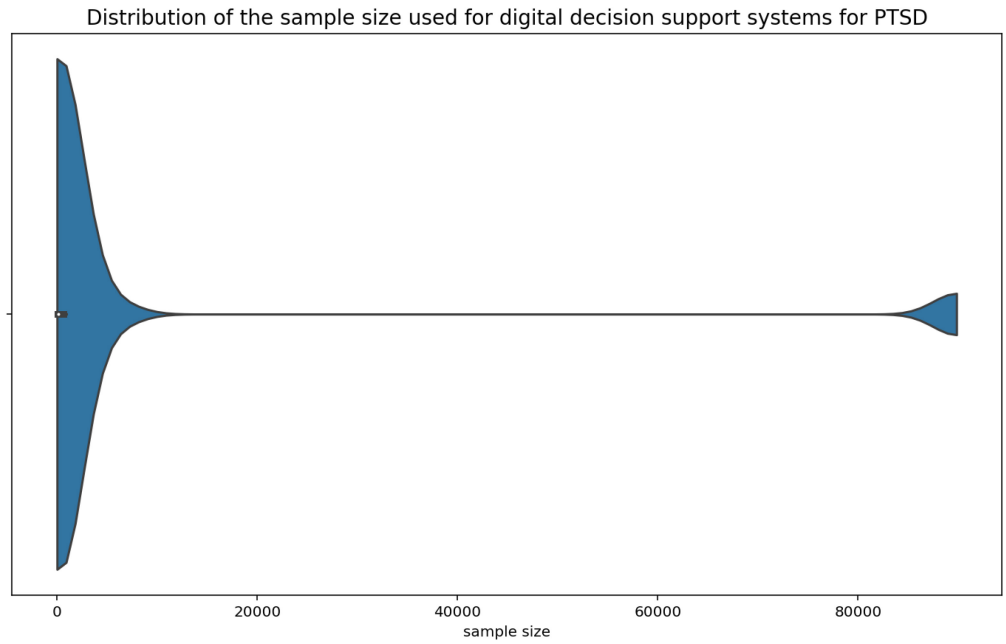


FIGURE 4
Sample size distribution.

papers investigated the diagnosis of PTSD (37, 41, 52, 60) and five focused on monitoring PTSD (33, 35, 56, 58, 59).

Data synthesis maturity level (EQ7)

The decision support systems were ranked according to the maturity scale described in see section "a framework for digital decision support systems." As stated by answering research question two, the majority of papers work with secondary data. This is supported by the high volume of research with a maturity level of two. **Figure 8** shows the number of articles grouped by maturity level.

Discussion

This research highlights the state of the art in digital decision support systems for PTSD based on our proposed framework. We developed the framework to ensure a holistic overview of all features of a DDSS. The dimensions of the framework represent the topics of interest and the choice of features is based on the conceptualization of the terminology extracted from the included articles dimension by dimension.

Concerning the data dimension, we noticed that questionnaires and checklists are still the most common and most mature (see **Figure 6**) input for decision support systems. When examining clinical guidelines like NICE (64) for diagnosing PTSD, questionnaires and checklists are still the only approach mentioned for diagnostics. Even though some new technologies, such as virtual or augmented reality, were investigated in the research found in this review, we noticed an absence of input parameters based on smartphones or wearables like GPS sensors or accelerometers. We hypothesize that this is due to the short life cycle of modern technologies, making it difficult to offer clinical evidence of their benefits. Questionnaires and checklists, however, have been around for many years and the methodology for administering them has not changed, therefore there is more scientific evidence of their use. Researchers and medical professionals are more likely to research, invest and adopt technology with strong evidence. This could be another reason why DDSSs using new technology are not widely included in clinical processes.

The data dimension also showed that the sample size is on average small and the statistical significance of the results was not proven by the majority of the research articles. Several reasons contribute to this. In general, medical data are hard to obtain for research because secondary use is still not easy with

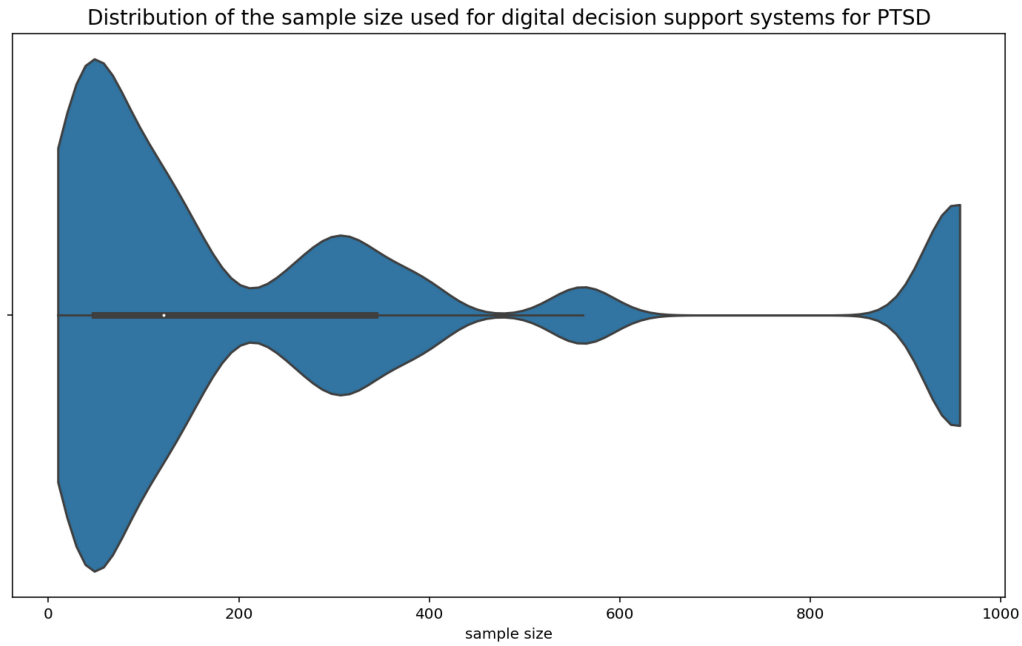


FIGURE 5
Sample size distribution excluding outliers.

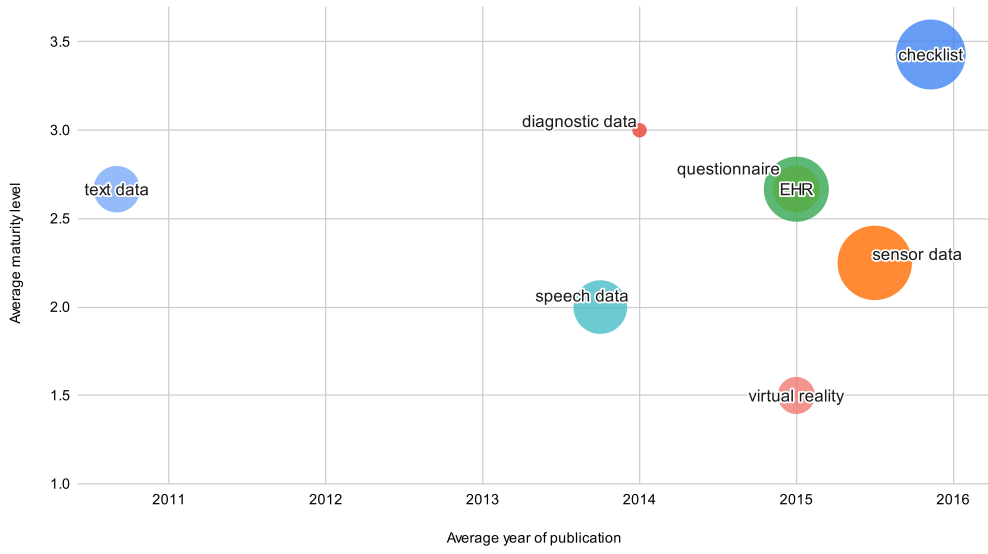


FIGURE 6
Data dimension concepts.

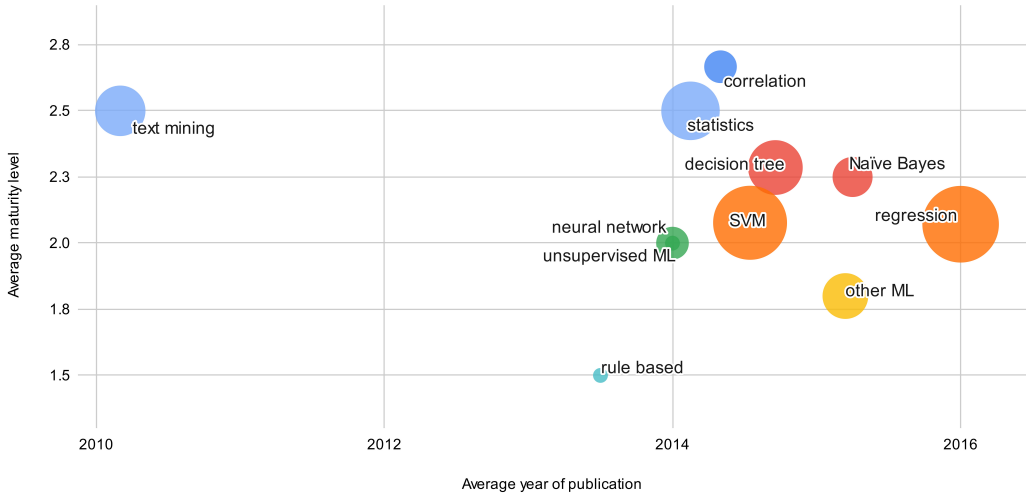


FIGURE 7
Decision technology concepts.

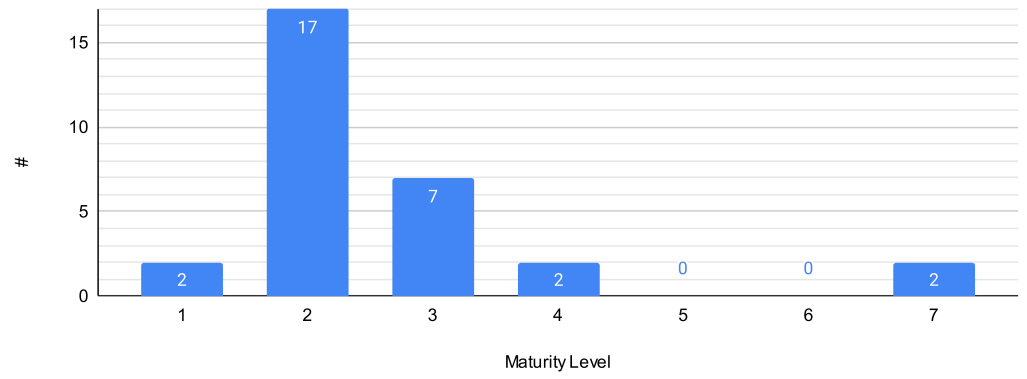


FIGURE 8
Bar chart – maturity levels.

many digital healthcare records and/or applications. Even if data can be obtained, they need to include the right parameters and have a structure that is usable for AI algorithms. Unstructured and text-based information is especially challenging to use for an AI. Further, most available datasets like the Jerusalem Trauma Outreach and Prevention Study do not include data on modern sensors (65).

The most common AI algorithm found during this literature review was support vector machines. Over the last few years, they have been developed to a *de facto* standard because they are easy to use, have good library support for programming and have low assumptions on the training data. We also observed that the number of research items resulting in usable products (maturity level $\geq$ 4) was low in three articles. Clinical studies with patient intervention (maturity level $\geq$ 3) were relatively low in nine papers out of 30. One reason for this could be that the small sample size of the research items does not provide sufficient evidence for clinical use.

All articles with a maturity level of 4 or more had, as one focus, validation of user acceptance and clearly defined user groups. Most articles with lower maturity levels did not have defined user groups. This could indicate a lack of strategic development and difficulties in bringing the research to a clinical setting. Our hypothesis is that interaction with

users or integration into clinical processes is often much more challenging to solve than intelligence of cognition. Still, most papers focus on cognition, not user interaction; our framework's validation dimension is evidence of this. We found 23 papers evaluating accuracy, which is an evaluation of AI technology, and five papers evaluating user acceptance or efficacy, meaning that they attempted to improve the current clinical process. Since most papers in our review are of maturity levels 1, 2 or 3 (meaning algorithm research), they do not include the clinical component necessary for user acceptance and efficacy evaluation. This shows a research gap when it comes to the enrichment of clinical processes with IT. The same goes for evaluating legal and IT-security constraints, which were not mentioned by any paper in our review. Since eHealth systems are getting increasingly focused by cyber attacks (66), IT and data security need to be a vital part of the evaluation to allow a safe DDSS adoption.

Further research has to be conducted on how the clinical process needs to be adapted for DDSSs to work, also in the context of the supported decisions. Most DDSS designers do not really understand the medical decision process but provide decisions in an "IT way." One limitation of this general hypothesis is that our research focuses solely on DDSS for PTSD. However, the narrow approach to include only PTSD shows that even in a very well-scoped area, a DDSS is hard to implement.

Since we used an inductive research approach to design our framework based on currently available literature, some important framework dimensions might be missing. One example is that the framework includes many technical aspects of the implementations and fewer organizational and financial perspectives. We encourage further research to include dimensions that describe the adoption of DDSSs in clinical processes.

Introducing our novel framework for DDSS, we provide a guide for decision support system evaluation. The framework is complementary to other healthcare technology evaluation methods (clinical, organizational, financial) and thus supports the design of comprehensive evaluation systems for DDSSs. Applying the maturity dimension helped us to examine what features of a DDSS are present, thereby indicating the steps to take in order to move up in maturity when developing decision support systems. Since the framework was developed out of general considerations, it can be applied to decision support systems outside of PTSD or mental health. However, it should be further evaluated to examine whether the terminology suits other domains. Higher maturity scales in particular need additional verification, since only two papers in our review had a maturity level above 4.

Conclusion

Our research aimed to analyze existing decision support systems for PTSD. Based on this goal, we developed a generic framework covering all dimensions of digital decision support systems. Our framework not only accelerates the development and benchmarking of DDSSs, but also acts as the foundation for our systematic literature review. Extracting data for all framework dimensions ensures consistency in our analysis and gives a holistic overview of DDSSs. During our review, we found working DDSS prototypes for PTSD and described their components. However, most of the systems are not evaluated in production use; they are only algorithmic models based on secondary datasets. This shows that there is still a gap between technical possibilities and actual clinical work. We proposed some possible explanations: small sample size, missing domain expertise, lack of focus to bring research to production. However, this gap should be analyzed further by testing our hypothesis and examining it with data from research on DDSSs for other mental diseases. For now, we conclude that only a few rare DDSSs for PTSD are ready for large-scale adoption in healthcare. The long-promised revolution of AI and ML for diagnosis in psychiatry, at least for PTSD, is yet to come.

Data availability statement

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

Author contributions

MB: conceptualization, methodology, investigation, resources, data curation, and writing – original draft. JM: investigation, data curation, and writing – original draft. PR: writing, review, editing, and supervision. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

References

1. Sauter VL. *Decision support systems for business intelligence*. Hoboken, NJ: John Wiley & Sons (1997). p. 618.

2. Power DJ. Decision support systems: A historical overview. In: Burstein FW, Holsapple C editors. *Handbook on decision support systems 1: Basic themes*. (Berlin: Springer) (2008). p. 121–40. doi: 10.1007/978-3-540-48713-5_7

3. Raymond RC. Use of the time-sharing computer in business planning and budgeting. *Manag Sci.* (1966) 12:B–363. doi: 10.1287/mnsc.12.8.B363

4. Turban E. The use of mathematical models in plant maintenance decision making. *Manag Sci.* (1967) 13:B–342. doi: 10.1287/mnsc.13.6.B342

5. Holt CC, Huber GP. A computer aided approach to employment service placement and counseling. *Manag Sci.* (1969) 15:573–94. doi: 10.1287/mnsc.15.11.573

6. American Psychiatric Association. *What is PTSD?* Virginia, VA: American Psychiatric Association (2020).

7. Spottswood M, Davydow DS, Huang H. The prevalence of posttraumatic stress disorder in primary care: A systematic review. *Harv Rev Psychiatry.* (2017) 25:159–69. doi: 10.1097/HRP.0000000000000136

8. Meltzer EC, Averbuch T, Samet JH, Saitz R, Jabbar K, Lloyd-Travaglini C, et al. Discrepancy in diagnosis and treatment of post-traumatic stress disorder (PTSD): Treatment for the wrong reason. *J Behav Health Serv Res.* (2012) 39:190–201. doi: 10.1007/s11414-011-9263-x

9. Sen A, Banerjee A, Sinha AP, Bansal M. Clinical decision support: Converging toward an integrated architecture. *J Biomed Inform.* (2012) 45:1009–17. doi: 10.1016/j.jbi.2012.07.001

10. Magrabi F, Ammenwerth E, McNair JB, Keizer NFD, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications. *Yearb Med Inform.* (2019) 28:128–34. doi: 10.1055/s-0039-1677903

11. Manogaran G, Lopez D. A survey of big data architectures and machine learning algorithms in healthcare. *Int J Biomed Eng Technol.* (2017) 25:182–211. doi: 10.1504/IJBET.2017.087722

12. Liang Z, Zhang G, Huang JX, Hu QV. Deep learning for healthcare decision making with EMRs. In: *Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Belfast: (2014). p. 556–9. doi: 10.1109/BIBM.2014.6999219

13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. *Brief Bioinform.* (2018) 19:1236–46. doi: 10.1093/bib/bbx044

14. Buchlak QD, Esmaili N, Leveque JC, Farrokhi F, Bennett C, Piccardi M, et al. Machine learning applications to clinical decision support in neurosurgery: An artificial intelligence augmented systematic review. *Neurosurg Rev.* (2019) 43:1235–53. doi: 10.1007/s10143-019-01163-8

15. Mettler T, Rohner P, Winter R. Towards a classification of maturity models in information systems. In: D'Atri A, De Marco M, Braccini AM, Cabiddu F editors. *Management of the interconnected world*. (Heidelberg: Physica-Verlag) (2010). p. 333–40. doi: 10.1007/978-3-7908-2404-9_39

16. Carvalho JV, Rocha Á, Abreu A. HISMM – Hospital information system maturity model: A synthesis. In: Mejia J, Muñoz M, Rocha Á, San Feliu T, Peña A editors. *Trends and applications in software engineering*. (Cham: Springer International Publishing) (2017). p. 189–200. doi: 10.1007/978-3-319-48523-2_18

17. Gomes J, Romão M. Information system maturity models in healthcare. *J Med Syst.* (2018) 42:235. doi: 10.1007/s10916-018-1097-0

18. Lichtenthaler U. Five maturity levels of managing AI: From isolated ignorance to integrated intelligence. *J Innov Manag.* (2020) 8:39–50. doi: 10.24840/2183-0606_008.001_0005

19. Ellefsen APT, Oleśków-Szłapka J, Pawłowski G, Toboła A. Striving for excellence in AI implementation: AI maturity model framework and preliminary research results. *LogForum.* (2019) 15:363–76. doi: 10.17270/J.LOG.2019.354

20. Saari L, Kuusisto O, Pirttikangas S. *AI maturity web tool helps organisations proceed with AI*. (2019). Available online at: https://cris.vtt.fi/en/publications/ai-maturity-web-tool-helps-organisations-proceed-with-ai (accesed June 19, 2020).

21. Loya SR, Kawamoto K, Chatwin C, Huser V. Service oriented architecture for clinical decision support: A systematic review and future directions. *J Med Syst.* (2014) 38:140. doi: 10.1007/s10916-014-0140-z

22. El-Sappagh SH, El-Masri S. A distributed clinical decision support system architecture. *J King Saud Univers Comput Inform Sci.* (2014) 26:69–78. doi: 10.1016/j.jksuci.2013.03.005

23. Basilakis J, Lovell NH, Celler BG. A decision support architecture for telecare patient management of chronic and complex disease. In: *Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. (Manhattan, NY: IEEE) (2007). doi: 10.1109/IEMBS.2007.4353296

24. Welch BM, Loya SR, Eilbeck K, Kawamoto KA. Proposed clinical decision support architecture capable of supporting whole genome sequence information. *J Pers Med.* (2014) 4:176–99. doi: 10.3390/jpm4020176

25. Lin HC, Wu HC, Chang CH, Li TC, Liang WM, Wang JYW. Development of a real-time clinical decision support system upon the web mvc-based architecture for prostate cancer treatment. *BMC Med Inform Decis Mak.* (2011) 11:16. doi: 10.1186/1472-6947-11-16

26. Sauser B, Verma D, Ramirez-Marquez J, Gove R. From TRL to SRL: The concept of systems readiness levels. In: *Proceedings of the Conference on Systems Engineering Research*. Los Angeles, CA: (2006).

27. Dataversity. *A pragmatic AI maturity model*. (2020). Available online at: https://www.slideshare.net/Dataversity/a-pragmatic-ai-maturity-model (accessed May 21, 2020).

28. Kitchenham B, Charters S. *Guidelines for performing systematic literature reviews in software engineering*. (2007). Available online at: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf (accessed June 30, 2022).

29. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of pubmed, scopus, web of science, and google scholar: Strengths and weaknesses. *FASEB J.* (2008) 22:338–42. doi: 10.1096/fj.07-9492LSF

30. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol.* (2006) 3:77–101. doi: 10.1191/1478088706qp063oa

31. Ma S, Galatzer-Levy IR, Wang X, Fenyö D, Shalev AY. A first step towards a clinical decision support system for post-traumatic stress disorders. *AMIA Annu Symp Proc.* (2016) 2016:837–43.

32. Barish G, Aralis H, Elbogen E, Lester P. A mobile app for patients and those who care about them: A case study for veterans with PTSD + anger. In: *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. (Trento: Association for Computing Machinery) (2019). p. 1–10. doi: 10.1145/3329189.3329248

33. Mallol-Ragolta A, Dhamija S, Boult TE. A multimodal approach for predicting changes in PTSD symptom severity. In: *ICMI – Proceedings of the International Conference Multimodal Interact*. (New York, NY: Association for Computing Machinery, Inc) (2018). p. 324–33. doi: 10.1145/3242969.3242981

34. Dabek F, Caban JJ. *A neural network based model for predicting psychological conditions*. (2015). Available online at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84945954195&doi=10.1007%2f978-3-319-23344-4_25&partnerID=40&md5=46be7a873337815b5db9d904941d4f1c (accessed June 30, 2022).

35. Xu R, Mei G, Zhang G, Gao P, Judkins T, Cannizzaro M, et al. A voice-based automated system for PTSD screening and monitoring. *Stud Health Technol Inform.* (2012) 173:552–8.

36. McWhorter J, Brown L, Khansa L. A wearable health monitoring system for posttraumatic stress disorder. *Biol Inspired Cogn Archit.* (2017) 22:44–50. doi: 10.1016/j.bica.2017.09.004

37. Omurca SI, Ekinci E. An alternative evaluation of post traumatic stress disorder with machine learning methods. In: *Proceedings of the INISTA – International Symposium on Innovations in Intelligent SysTems and Applications*. (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc) (2015). doi: 10.1109/INISTA.2015.7276754

38. He Q, Veldkamp BP, Glas CAW, de Vries T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment.* (2017) 24:157–72. doi: 10.1177/1073191115602551

39. Myers CE, Radell ML, Shind C, Ebanks-Williams Y, Beck KD, Gilbertson MW. Beyond symptom self-report: Use of a computer "avatar" to assess post-traumatic stress disorder (PTSD) symptoms. *Stress.* (2016) 19:593–8. doi: 10.1080/10253890.2016.1232385

40. Karstoft KI, Galatzer-Levy IR, Statnikov A, Li Z, Shalev AY, Ankri Y, et al. Bridging a translational gap: Using machine learning to improve the prediction of PTSD. *BMC Psychiatry.* (2015) 15:30. doi: 10.1186/s12888-015-0399-8

41. Finkelman MD, Lowe SR, Kim W, Gruebner O, Smits N, Galea S. Customized computer-based administration of the PCL-5 for the efficient assessment of PTSD: A proof-of-principle study. *Psychol Trauma.* (2017) 9:379–89. doi: 10.1037/tra0000226

42. Karstoft KI, Statnikov A, Andersen SB, Madsen T, Galatzer-Levy IR. Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers. *J Affect Disord.* (2015) 184:170–5. doi: 10.1016/j.jad.2015.05.057

43. Miner A, Kuhn E, Hoffman JE, Owen JE, Ruzek JI, Taylor CB. Feasibility, acceptability, and potential efficacy of the PTSD coach app: A pilot randomized controlled trial with community trauma survivors. *Psychol Trauma.* (2016) 8:384–92. doi: 10.1037/tra0000092

44. Pyne JM, Constans JI, Wiederhold MD, Gibson DP, Kimbrell T, Kramer TL, et al. Heart rate variability: Pre-deployment predictor of post-deployment PTSD symptoms. *Biol Psychol.* (2016) 121:91–8. doi: 10.1016/j.biopsycho.2016.10.008

45. Zhuang X, Rozgić V, Crystal M, Marx BP. Improving speech-based PTSD detection via multi-view learning. In: *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT)*. (Manhattan, NY: IEEE) (2014). p. 260–5. doi: 10.1109/SLT.2014.7078584

46. Scherer S, Stratou G, Gratch J, Morency LP. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In: *Proceedings of the Annual Conference International Speech Communication Association*. Lyon: (2013). p. 847–51. doi: 10.21437/Interspeech.2013-240

47. Dabek F, Caban JJ. Leveraging big data to model the likelihood of developing psychological conditions after a concussion. In: Roy A, Venayagamoorthy K, Alimi A, Angelov P, Trafalis T editors. *Procedia computer science*. (Amsterdam: Elsevier) (2015). p. 265–73. doi: 10.1016/j.procs.2015.07.303

48. Alvarez-Conrad J, Zoellner LA, Foa EB. Linguistic predictors of trauma pathology and physical health. *Appl Cognit Psychol.* (2001) 15:S159–70. doi: 10.1002/acp.839

49. Saxe GN, Ma S, Ren J, Aliferis C. Machine learning methods to predict child posttraumatic stress: A proof of concept study. *BMC Psychiatry.* (2017) 17:223. doi: 10.1186/s12888-017-1384-1

50. Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in twitter. In: *Proceedings of the International Conference Weblogs Social Media*. (Palo Alto, CA: The AAAI Press) (2014). p. 579–82.

51. Ćosić K, Popović S, Kukolja D, Horvat M, Dropuljić B. Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders. *Cyberpsychol Behav Soc Netw.* (2010) 13:73–8. doi: 10.1089/cyber.2009.0260

52. Marinić I, Supek F, Kovačić Z, Rukavina L, Jendričko T, Kozarić-Kovačić D. Posttraumatic stress disorder: Diagnostic data analysis by data mining methodology. *Croat Med J.* (2007) 48:185–97.

53. Kuhn E, Greene C, Hoffman J, Nguyen T, Wald L, Schmidt J, et al. Preliminary evaluation of PTSD coach, a smartphone app for post-traumatic stress symptoms. *Mil Med.* (2014) 179:12–8. doi: 10.7205/MILMED-D-13-00271

54. Galatzer-Levy IR, Karstoft KI, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: A machine learning application. *J Psychiatr Res.* (2014) 59:68–76. doi: 10.1016/j.jpsychires.2014.08.017

55. Scherer S, Lucas GM, Gratch J, Rizzo A, Morency LP. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Trans Affect Comput.* (2016) 7:59–73. doi: 10.1109/TAFFC.2015.2440264

56. Zatzick D, O'Connor SS, Russo J, Wang J, Bush N, Love J, et al. Technology-enhanced stepped collaborative care targeting posttraumatic stress disorder and comorbidity after injury: A randomized controlled trial. *J Trauma Stress.* (2015) 28:391–400. doi: 10.1002/jts.22041

57. Shaikh al arab A, Guédon-Moreau L, Ducrocq F, Molenda S, Duhem S, Salleron J, et al. Temporal analysis of heart rate variability as a predictor of post traumatic stress disorder in road traffic accidents survivors. *J Psychiatr Res.* (2012) 46:790–6. doi: 10.1016/j.jpsychires.2012.02.006

58. Freeman D, Antley A, Ehlers A, Dunn G, Thompson C, Vorontsova N, et al. The use of immersive virtual reality (VR) to predict the occurrence 6 months later of paranoid thinking and posttraumatic stress symptoms assessed by self-report and interviewer methods: A study of individuals who have been physically assaulted. *Psychol Assess.* (2014) 26:841–7. doi: 10.1037/a0036240

59. Hossain MF, George O, Johnson N, Madiraju P, Flower M, Franco Z, et al. editors. Towards clinical decision support for veteran mental health crisis events using tree algorithm. In: Getov V, Gaudiot J-L, Yamai N, Cimato S, Chang M, Teranishi Y, et al. editors. In: *Proceedings of the International Computer Software and Applications Conference*. (Washington, DC: IEEE Computer Society) (2019). p. 386–90. doi: 10.1109/COMPSAC.2019.10237

60. Gong Q, Li L, Tognin S, Wu Q, Pettersson-Yeo W, Lui S, et al. Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level. *Psychol Med.* (2014) 44:195–203. doi: 10.1017/S0033291713000561

61. World Health Organization. *ICD-10 : International statistical classification of diseases and related health problems : Tenth revision*. Geneva: World Health Organization (2004).

62. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. 5th ed. Virginia, VA: American Psychiatric Association (2013). doi: 10.1176/appi.books.9780890425596

63. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* (1997) 30:1145–59. doi: 10.1016/S0031-3203(96)00142-2

64. National Institute for Health and Care Excellence. *Post-traumatic stress disorder – [D] evidence reviews for psychological, psychosocial and other non-pharmacological interventions for the treatment of PTSD in adults*. London: National Institute for Health and Care Excellence (2020).

65. Shalev AY, Ankri Y, Israeli-Shalev Y, Peleg T, Adessky R, Freedman S. Prevention of posttraumatic stress disorder by early treatment: Results from the Jerusalem trauma outreach and prevention study. *Arch Gen Psychiatry.* (2012) 69:166–76. doi: 10.1001/archgenpsychiatry.2011.127

66. Bertl M. News analysis for the detection of cyber security issues in digital healthcare. *Young Inform Sci.* (2019) 4:1–15.

Appendix 2

[II]

M. Bertl, P. Ross, and D. Draheim. A Survey on AI and Decision Support Systems in Psychiatry – Uncovering a Dilemma. *Expert Systems with Applications*, 202:117464, 2022

Expert Systems With Applications 202 (2022) 117464

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A survey on AI and decision support systems in psychiatry – Uncovering a dilemma

Markus Bertl [a,*], Peeter Ross [a], Dirk Draheim [b]

[a] Department of Health Technologies, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
[b] Department of Software Science, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

ARTICLE INFO

Keywords:
Medical information policy
Medical technology
Digital decision support system (DDSS)
Clinical Decision Support Systems (CDSS)
Artificial Intelligence (AI)
Machine Learning (ML)
Psychiatry

ABSTRACT

Every year, healthcare specialists collect more and more data about patients but struggle to use it to optimize disease prevention, diagnosis, or treatment processes. While a manual use of this medical data is virtually impossible considering the vast growth rate, automation with artificial intelligence (AI) and digital decision support systems (DDSSs) has still not yielded any large-scale success in healthcare. We aim to investigate possible obstacles, the trustworthiness based on potential biases, and the adoption of new technology by AI and DDSSs in psychiatry based on a systematic literature review. We screened 520 papers about AI or DDSSs in psychiatry. We added results from a literature screening of 65 articles about AI or DDSSs for post-traumatic stress disorder as one specific psychiatric disease to our research, given that literature possibly deviates from general decision support systems for psychiatry. Out of 80 articles, we extract algorithms, data collection method and sample size of the used training data, and testing process including accuracy metrics. The results show that sample sizes are small (median of 151.5), a focus on algorithm development without real-world interaction, and methodological shortcomings when it comes to the evaluation of DDSSs. Our survey concludes that DDSSs in psychiatry are not ready for the often-promised "AI revolution in healthcare".

1. Introduction

Health data is growing steadily. According to an estimation by the International Data Corporation (IDC) 2.414 exabytes of health data were generated by the end of 2020. Given that time is scarce, it is already impossible to read all medical data of a patient before a doctor's appointment, stay up-to-date with treatment methods, or track drug-drug interaction manually. Especially in psychiatry, the "gold standard" of human diagnosis has low accuracy (Aboraya et al., 2006; Al-Huthail, 2008; AlSalem et al., 2020; Hamidia et al., 2022; Kitamura et al., 1989). McGlynn et al. (2003) suggest that more than half of patient care in the U.S. is not administered according to medical guidelines. Furthermore, 52.7% of people with depression are not correctly diagnosed by their general practitioner (Mitchell et al., 2009). Low accuracy of initial psychiatric diagnoses has also been reported by AlSalem et al. (2020). This evidence shows that we collect medical data but struggle to make use out of it. Digital Decision Support Systems (DDSSs) and artificial intelligence (AI) could be one way to address diagnostic uncertainty by assisting medical professionals in making sense of data. In this research, we understand DDSSs as defined by Sauter, 1997 as "computer-based systems that bring together information from a variety of sources, assist in the organization and analysis of information and facilitate the evaluation of assumptions underlying the use of specific models" (Sauter, 1997). Artificial Intelligence is defined by the Cambridge Dictionary as "the study of how to produce computers that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, solve problems, and learn".[1]

Mounting evidence suggests that demand for such systems is given. 56% of U.S. adults are willing to share their health data with tech companies like Google (Day et al., 2019). The big data market for health data is booming (Dash et al., 2019) and is estimated to reach 7 billion USD by 2021. However, considerable doubt exists. 85.9% of office-based physicians use electronic health records in U.S. (Office-Based Physician Electronic Health Record Adoption, 2019). A similar trend can be observed in the European Union (eHealth, Well-being, and Ageing (Unit H.3), 2019). In today's digitized world, even a single byte can have importance in health-related decisions. Digitized data shows what kind of treatment a person gets, what kind of medication is prescribed,

* Corresponding author.
E-mail addresses: mbertl@taltech.ee (M. Bertl), peeter.ross@taltech.ee (P. Ross), dirk.draheim@taltech.ee (D. Draheim).

[1] dictionary.cambridge.org/dictionary/english/artificial-intelligence

https://doi.org/10.1016/j.eswa.2022.117464
Received 23 January 2021; Received in revised form 19 April 2022; Accepted 28 April 2022
Available online 2 May 2022

whether a person is allowed to drive, or even whether someone is allowed to make decisions on his own. Security of these data is not guaranteed, breaches have happened, data has been manipulated, and health IT has even been targeted by terroristic activities (Bertl, 2019). These developments could be countered by increasing investments in security, data protection techniques, and zero-trust computing. But AI itself also introduces new challenges to clinical safety (Challen et al., 2019). For example, people increase their worry about whether AI can be trusted, if possible reasons for bias have been taken into account, if DDSSs are tested enough to be used in such a sensitive area as health, and what accuracy we can expect.

A wide variety of biases in scientific publications have been studied extensively. Most notably, scholars have famously argued that false discovery rates of what researchers advertise as "experimental research findings" in scientific publications exceeds 50% (Ioannidis, 2005).

In DDSS research, some researchers also begin to critique the datasets used for AI and the dataset culture in machine learning (Paullada et al., 2020). Since data is the foundational part of AI and machine learning, the question arises whether currently used data is curated well enough. Furthermore, scholars increasingly cast doubt on whether sampling methodologies are good enough for justifying their use in the medical domain. Poorly curated datasets reflect human biases. Given their foundational role for computerized systems, human biases run the risk of spreading flawed decisions at a large-scale, possibly with catastrophic consequences. These arguments lead to the question of what the current state of the art concerning AI and decision support systems is.

The facts around the dilemma of growing data, medical complexity, and trustworthiness show that a thorough investigation of DDSSs and AI in healthcare needs to be conducted. In this paper, we investigate the corresponding literature to find obstacles, indications about the trustworthiness, and the use of emerging technologies of AI and DDSSs in psychiatry. Our goal is to examine the state of the art and possible ways of DDSS improvement.

2. Medical background

Psychiatric disorders represent critical non-communicable diseases of the 21st century. In 2010, mental disorders accounted for €461 billion in healthcare costs in Europe (Gustavsson et al., 2011) and ranked as the leading cause of years lived with disabilities (Wittchen et al., 2011). However, diagnostic accuracy in psychiatry is still low. For example, 69% of patients with bipolar disorder are initially misdiagnosed by mental health specialists (Singh & Rajput, 2006). Such errors in diagnoses remain uncorrected for an average of 5.7 years (Morselli & Elgie, 2003). Despite some achievements in the implementation of DDSSs in clinical routines like drug-drug interaction databases (Metsallik et al., 2018), primary care or hospital DDSSs (Sutton et al., 2020), medical specialists have been waiting for a breakthrough of AI-based DDSSs in healthcare settings for at least two decades without tangible success. Most notably, systems still suffer from both low user acceptance and adoption rates (Bates et al., 2003; Gaube et al., 2021; Sittig et al., 2006). While new medical devices supported by software, like diagnostic devices, digital imaging, or cardiovascular interventional equipment, incorporate new technology (Bettinger, 2018; Neuman et al., 2012), are well-accepted by clinicians and have quickly acquired substantial market shares (Schreyögg et al., 2009), DDSSs have not followed a similar trajectory. Innovative technology like AI does mostly not deliver value in clinically adopted DDSSs (Strickland, 2019). This may be the case because medical devices are well-targeted at specific clinical professionals and lead to better performance, while DDSS developers strive to cover a wide range of clinical disciplines with one technological application. These tendencies exacerbate for more specialized healthcare sections such as psychiatry. More precisely, the corresponding situation in psychiatry differs from other medical domains because biomarkers and technical tools for decision-making have not yet been validated. As a result, diagnoses and treatment decisions tend to depend on clinical interviews, observations, and self-report measures (Maron et al., 2019). These do currently not deliver as precise results as biomarkers do. Despite the urgent need caused by increasing data, increasing medical complexity, as well as limited staff and financial capacities in healthcare, DDSSs and AI still suffer a niche existence. Software is still mainly used to store data rather than as a tool to redesign care processes or improve decision quality and safety. Therefore, investigating different aspects that might hinder the broader adoption of DDSSs in medicine is of great interest among clinicians. We contribute to this debate by analyzing current obstacles and ways to improve AI-based DDSSs in psychiatry.

Additionally to psychiatry as a whole, post-traumatic stress disorder (PTSD) was taken as a clinical entry in psychiatry. The American Psychiatric Association defines PTSD as "a psychiatric disorder that can occur in people who have experienced or witnessed a traumatic event such as a natural disaster, a serious accident, a terrorist act, war/combat, rape or other violent personal assault"[2]. People with PTSD experience recurrent thoughts about their traumatic experience which influences their daily life. The lifetime prevalence of PTSD is around 12.5% (Spottswood et al., 2017), which renders it all the more pressing to examine this disorder in greater depth. Even more so, people suffering from PTSD are often un- or misdiagnosed, resulting in wrong, incomplete, or missing treatment (Meltzer et al., 2012).

3. Methods

3.1. Reporting Standards

We follow Kitchenham & Charters' (2007) five stages for performing systematic literature reviews in software engineering:

(1) Search Strategy
(2) Study Selection
(3) Study Quality Assessment
(4) Data Extraction
(5) Data Synthesis

The process of conducting this literature review is visualized in Fig. 1. Importantly, our research methodology complies with the PRISMA checklist for transparent reporting of systematic reviews and *meta*-analyses (Liberati et al., 2009; Moher et al., 2009).

We want to highlight that our study is a systematic literature review. Doing a *meta*-analysis is not possible because most AI/ML research does not report effect sizes with confidence intervals which would then be used in a fixed or random effect model for synthesis. Because of that, we do not assign weights to the extracted features of the studies based on their sample sizes.

3.2. Research questions

For the literature search, we worked based on RQ1, RQ2, and RQ3 shown in Table 1. The results of our survey were then used to answer RQ4 based on a narrative synthesis.

3.3. Search strategy

We built a search string derived from the above research questions. This search string consists of the objects of interest (decision support or artificial intelligence or subcategories of machine learning) and the scope of our review (psychiatry). We restricted our search to articles in English with a publication date between 2000 and 2020 to only include modern technology. The resulting search string was (((decision AND support AND system) OR (artificial AND intelligence) OR ((machine OR

[2] https://psychiatry.org/patients-families/ptsd/what-is-ptsd

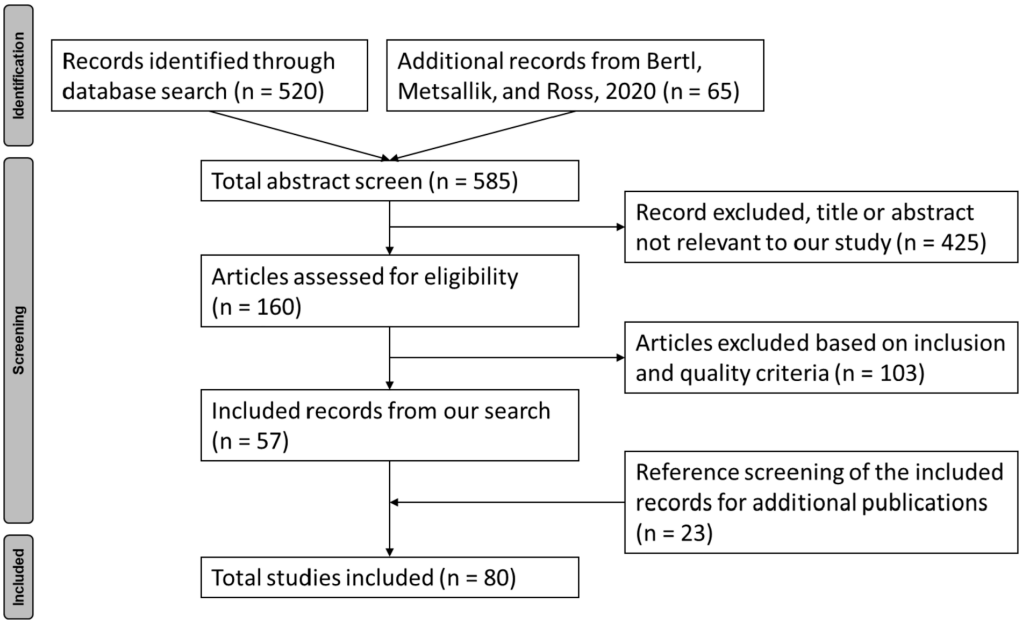


Fig. 1. search strategy for the literature review.

Table 1
research questions.

#	Research Question
RQ1	What are the current obstacles of research on AI and decision support systems in psychiatry?
RQ2	How trustworthy is the state of the art concerning AI and decision support systems in psychiatry?
RQ3	How do AI and decision support systems in psychiatry adopt new technology?
RQ4	What is needed to improve AI and decision support systems in psychiatry?

deep OR supervised OR unsupervised OR reinforcement) AND learning)) AND psychiatry). Additionally to psychiatry as a whole, we supplement our analysis on general DDSSs and AI algorithms for psychiatry by also selecting systems designed for one specific disease. By that, we are reducing possible bias since approaches for the whole domain of psychiatry might have different results than for one condition. For that, we used the data from Bertl et al. (2020), with a similar search string for Post Traumatic Stress Disorder (PTSD). We applied our search strings to the research papers' titles, abstracts, and tags in Scopus' abstract and citation database. Scopus was chosen as the primary source because it is the largest abstract and citation database of research literature with 100% MEDLINE coverage (Falagas et al., 2008).

Our Scopus search was carried out on 11th October 2021. We also conducted reference screening and a manual search in Google Scholar and the web to find additional research.

3.4. Study selection

Titles and abstracts of queried articles were analyzed to identify relevant articles derived from our search results. Articles that fitted the research questions and met the inclusion criteria (see 3.4.1), and the quality criteria (see 3.4.2) were included. To reduce bias, title and abstract screening, as well as checking the inclusion and quality criteria were conducted independently by two researchers. The two sets were then merged, and deviations were discussed among the authors. 11 articles were excluded because no full-text could be retrieved. In the end, we selected 49 papers from our psychiatry search and 30 papers from our PTSD search, a total of 80 articles for this review. Cohen's Kappa was calculated to assess interrater reliability (McHugh, 2012). The agreement score was 90% (Cohen's Kappa 0.718). All disagreements could be resolved and were mainly concerned with whether maturity level 1 or 2 studies were based on computerized or paper-based algorithms (IC2).

3.4.1. Inclusion criteria

Table 2 presents the inclusion criteria used for our literature search.

3.4.2. Study quality assessment

Since uncovering possible research biases was one purpose of this review, we reduced study quality assessments to a minimum to get a more holistic view of the published research. Quality criteria are shown in Table 3. We added QC3 since we found two articles originating from journals without peer review through reference search.

3.5. Data extraction and synthesis

To answer our research questions, clear scoped questions for data extraction were formed (see Table 4) based on the DDSS framework further described in Bertl et al. (2020). The framework was created based on thematic analysis (Braun & Clarke, 2006) to define different components of DDSSs in healthcare. We use the following dimensions for our extraction:

Table 2
inclusion criteria for our literature search.

#	Inclusion Criteria
IC1	Does the study deal with decision support systems (e.g. systems that help to diagnose, screen, predict, or treat)?
IC2	Does the study use computerized statistical, AI/ML, or rule-based algorithms?
IC3	Does the article deal with psychiatric diseases or psychiatric problems?
IC4	Is the article related to at least one of our research questions?

Table 3
quality criteria for our literature search.

#	Quality Criteria
QC1	Has the study a well-defined structure?
QC2	Does the study bring evidence for the proposed approach (either by citing relevant literature or validating the results)?
QC3	Has the research been peer-reviewed?

Table 4
extraction questions (EQ) for data collection from the literature.

#	Extraction parameters
EQ1	What data do existing decision support systems use?
EQ1.2	How large is the used sample size?
EQ2	How are existing DDSSs in mental health implemented?
EQ2.1	Decision technology
EQ2.2	User interaction technology
EQ2.3	Data collection technology
EQ3	Which features were validated?
EQ3.1	How high is accuracy?
EQ4	What medical domains/diseases are currently targeted by the DDSS?
EQ5	What decisions are supported by the system?
EQ6	What maturity level does the DDSS have?

- *Data* used by the DDSS to get insights into what data sources are used and what average sample sizes are. This contributes to RQ2 by investigating if data sources are trustworthy and sample sizes are appropriate. It also contributes to RQ3 by showing what technology is used for data collection.
- *Technology* for data collection, user interaction, and decision-making contributes to RQ3.
- *Validation* around accuracy, user acceptance, efficacy, and legal/compliance to extract information around RQ2 - trustworthiness.
- *Medical Domain* or diseases that DDSSs are applied to. Different diseases can be used for possible subgroup analysis.
- *Decision type* (prediction, diagnosis, screening, monitoring, or treatment) for possible subgroup analysis.
- *Maturity level* of DDSSs from 1 (idea of DDSS) to 7 (world-wide adopted product). This indicates general DDSS adoption rates by showing how they have advanced in the market. Our maturity levels contribute to RQ1 and RQ4.

Using a framework for data extraction helps us to make our research reproducible by highlighting which features of DDSSs were used to answer our research questions. We omitted the framework dimension "user group" during extraction as it does not contribute to our research questions.

We decided to reuse our own framework since it was developed especially for the analysis of research about DDSSs in psychiatry and has already been applied successfully. Other existing frameworks were found to be either too complex (Boza et al., 2009; Sprague, 1980) or not fit our research questions (Camacho et al., 2020; Sim & Berlin, 2003). Greenes et al. also highlight that too many different perspectives on studying DDSSs make a single DDSS model which can be reused for different applications challenging (Greenes et al., 2018).

The extracted answers to the EQ's were then combined into a feature matrix based on a common agreement among the authors. The extracted features were then clustered to have a common terminology that allows further analysis and the possibility to compare results based on a narrative synthesis. Fig. 1 highlights our search process.

3.6. Risk of bias

We used the funnel plots based on sample size and accuracy to search for possible publication biases (Sterne & Harbord, 2004). Funnel plots plot the treatment effect (accuracy in our case) against the sample size. Suppose studies with smaller sample sizes have equal or less variance than studies with higher sample sizes (accuracy's distribution is skewed). In that case, publication bias can be assumed (Kitchenham & Charters, 2007). Our empirical accuracy distribution does not indicate publication bias since it is nearly symmetric with only a small left-skew of -0.03. Studies with smaller sample sizes have more variance in accuracy than studies with higher sample sizes. However, these results should be interpreted with caution since not all studies mentioned their systems' accuracy values. Some studies used different metrics that could not be converted to accuracy (see 4.1.3). Since most articles did not yield statistically significant results, alternative methods for detecting publication bias or data mining like p-curve analysis were not possible. However, Fig. 2 shows that mostly articles with high accuracy scores have been published.

4. Results

4.1. Facts from the literature

This sub-section deals with the hard facts obtained from the selected literature based on the extraction questions in Table 4. The results are then analyzed, synthesized, and discussed in section 4.2. In general, article publication dates range from 2001 to 2020. About 65% of articles were published between 2014 and 2020 (51/80). The majority of articles were published in medical journals (50/80), 16 were published in computer science journals, and 14 in journals specific to digital healthcare or health informatics.

4.1.1. Data

As shown in Fig. 3, the majority of DDSSs uses the results of questionnaires or checklists (22.5%). As of the time of this review, innovative technology like virtual reality or sensors for digital phenotyping has not been adopted widely.

Fig. 4 shows the distribution of sample size. More precisely, sample size's mean was $\mu = 4133$ with a standard deviation of $\sigma = 16147$ and a median of $\eta = 151.5$. The smallest sample size observed was 4, the highest 89840. Outliers (5972, 11540, 14929, 45388, 89840, 89840) have been removed from the plot for better visibility. Additionally to total sample size, 33 articles listed information about the number of positive and negative cases in their datasets.

Out of the 80 articles in our review did not mention possible biases of data collection, their data, or the DDSS algorithm.

4.1.2. Technology

Fig. 5 shows the different algorithms used for DDSSs in psychiatry. 24 research projects used Support Vector Machines (SVM) as decision algorithm for their system. The second most popular decision technology was logistic regression (19 articles). Together with decision trees and random forests (12 articles), these groups make up nearly two-thirds of all algorithms. Explainability or explainable AI (XAI) was not mentioned by any of the papers in this review. Two papers mentioned that their approach is a black box as a downside.

Because of generally low maturity scores described in 4.1.5, no clear indication of user interaction or data collection technology could be found since these technologies are typically not present when only dealing with datasets.

4.1.3. Validation

Fig. 6 presents the evaluation methods of DDSSs. Since the majority of research found is based on algorithm development based on datasets, the most dominant evaluation criterion was algorithmic accuracy measured by precision, recall, F1 score, or area under the receiver operating characteristic (ROC) curve (AUC). Definitions of the mentioned performance measure, especially AUC, can be found at Bradley (1997) 80% of the articles used algorithmic accuracy for the evaluations.

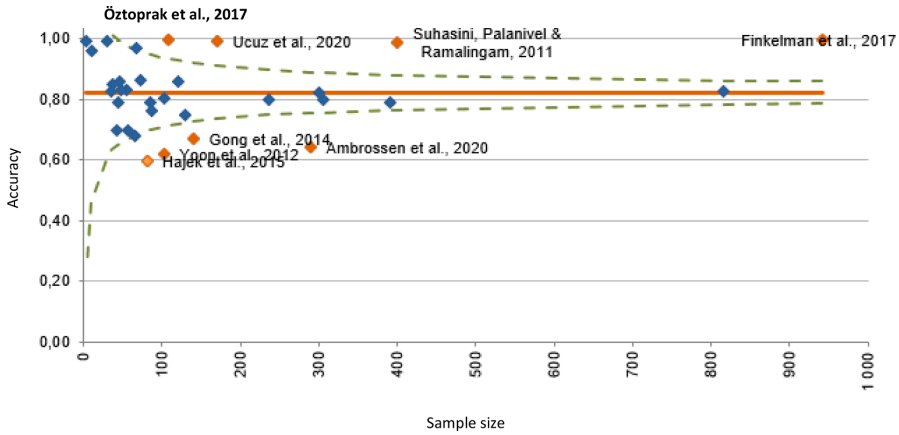


Fig. 2. funnel plot – accuracy vs. sample size.

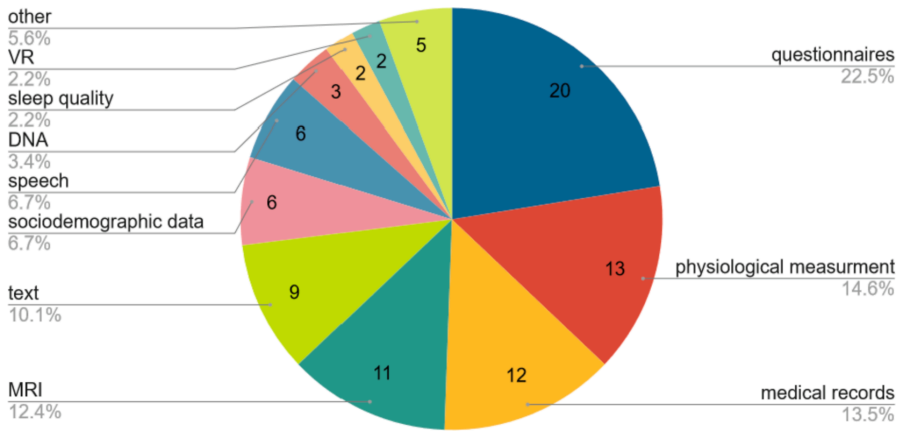


Fig. 3. input data types of DDSSs.

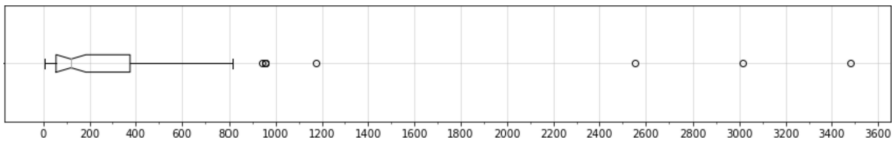


Fig. 4. distribution of the sample size of DDSSs with removed outliers.

Most popular was the measurement "accuracy", present in 34 out of the 80 studies. The second most popular performance measurement was the area under the curve, present in 20 studies. Since accuracy and AUC values cannot be converted to each other, we extracted all accuracy measurements present in each paper and aggregated each scale individually. Mean accuracy of the DDSSs is $\mu = 82.8\%$ with a median of $\eta = 82.5\%$ and a standard deviation of $\sigma = 0.116$ (Fig. 7). Mean area under the curve value is $\mu = 0.809$ with a median of $\eta = 0.805$ and a standard deviation of $\sigma = 0.071$. 42 papers listed a confusion matrix or precision/recall values apart from other evaluation metrics like AUC, F1, or accuracy scores.

Fig. 8 shows the accuracy scores with the corresponding sample size of the different papers in our analysis. For better visibility, we removed the two outliers with sample size 89840.

4.1.4. Supported Medical Domains/Diseases and Supported Decisions.

Table 5 lists the diseases which are currently supported by the DDSSs in psychiatry.

Extracted features for supported diseases and decisions did not yield any results that could be linked to our research questions about problems, biases, and fairness of AI algorithms.

4.1.5. Maturity

Based on the digital decision support framework described by Bertl et al. (2020), maturity was ranked on a scale from 1 (idea) to 7 (worldwide adopted product). The levels are described in Table 6. Scores based

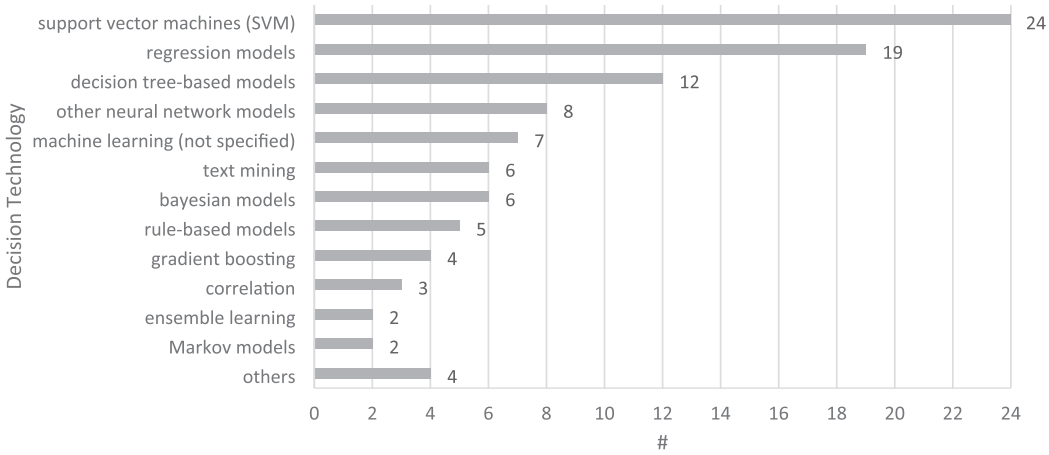


Fig. 5. decision technology used by DDSSs.

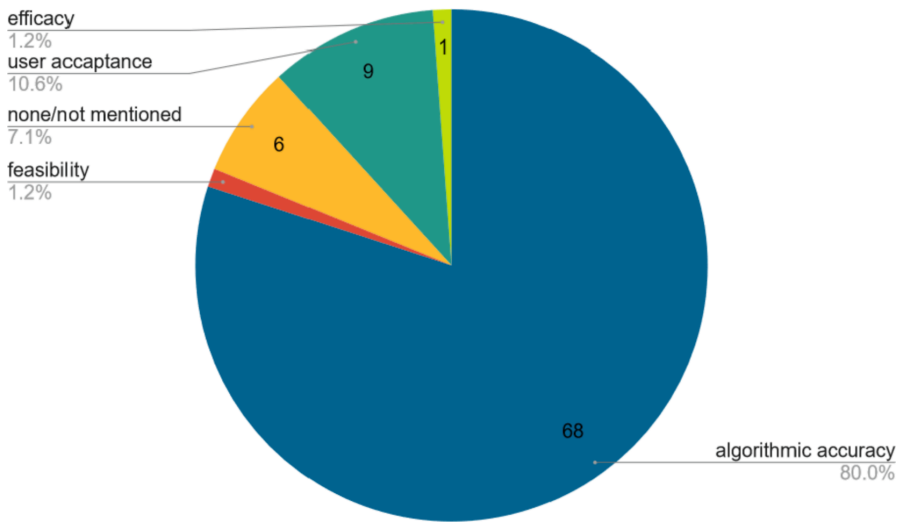


Fig. 6. evaluation methods of DDSSs.

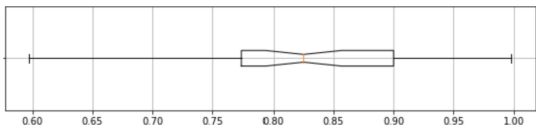


Fig. 7. distribution of the accuracy of DDSSs.

on these maturity levels indicate advancement in the development of AI and DDSSs in psychiatry.

The maturity levels of the research in this review are shown in Fig. 9. The majority of articles dealt with maturity levels two (30) and three (38). The average maturity level of research in this survey was 2.625 and therefore indicates that most research deals with algorithm development.

A Mann-Whitney-U test (McKnight & Najab, 2010) indicated that the

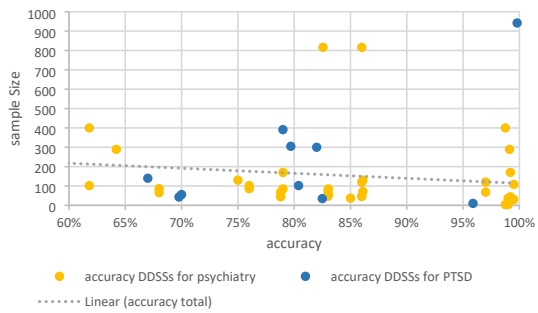


Fig. 8. scatter plot – sample size vs. accuracy.

Table 5
diseases supported by DDSSs in psychiatry.

Disease	#
depression	10
schizophrenia	8
psychotic disorder	3
anxiety	3
PTSD	3
bipolar disorder	2
ADHD	2
suicidality	2
others	5

Table 6
maturity level scale for DDSSs.

Level	Description
1	Idea without implementation
2	Implementation without real-world interaction (algorithm development)
3	Implementation with real-world interaction but without patient intervention (no real intervention on a patient takes part based on the output of the DDSS)
4	Fully functioning prototype, system triggers real-world action (e.g., clinical trial)
5	Operational product (at least one adopter, certified if required)
6	Locally adopted product
7	World-wide adopted product (transformational)

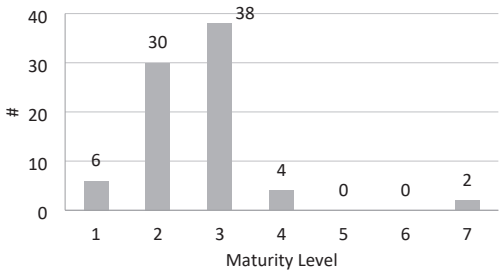


Fig. 9. maturity levels (according to Table 6) of DDSSs.

difference between maturity levels of articles about DDSSs and AI in psychiatry does not differ statistically significantly from maturity levels for articles about DDSSs and AI for PTSD ($U(n_{psychiatry} = 48, n_{PTSD} = 30) = 544.5$, $z = 1.79736$, $p < .07186$).

4.2. Findings & discussion

Our research summarizes the current state of the art on decision support systems and artificial intelligence in psychiatry based on a theoretical framework for DDSSs.

4.2.1. Current Obstacles of Research of AI and Decision Support Systems in Psychiatry – RQ1

Concerning RQ1 "obstacles", we observed small sample sizes when examining data dimensions. Often, either datasets that were already available were used, or data was collected using local medical information systems. No data source covering the whole population at the state or national level was found. We assume that this indicates a lack of standardized eHealth infrastructures, universal data access, interoperability, and a lack of data reuse capabilities in psychiatry. Psychiatry, in particular, is still mainly based on unconnected, impractical, or inefficient electronic record systems. Sometimes, documentation is even paper-based. This can introduce selection bias to AI training data. Clinical notes in EMR are mostly entered in free text. Coding and classification of findings are either based on very general, artificial categories like ICD or DSM, or locally used legacy taxonomies. It is questionable whether these artificial categories reflect the actual mental problem present with a patient. Knowing about the problems, new frameworks like the Research Domain Criteria (RDoC), which take into account more dimensions than just patient symptoms, are currently developed (Cuthbert, 2014). This is especially important given that we found a small correlation between sample size and accuracy, indicating that more data does not necessarily produce better outcomes. Instead, quality and representativeness of the data remain the important factor. As we have written at the beginning of this section, such problems continue to bedevil research on healthcare. Additionally to quality issues, fragmentation of healthcare data makes it difficult to successfully implement DDSSs and AI in real-world scenarios. This conclusion is shared by Panch et al. (2019). Although data is a fundamental part of AI success, it alone is not sufficient to solve all problems DDSSs are currently facing. Research's generally low maturity scores also indicate a problem of bringing AI and DDSSs into clinical practice. Besides health data's fragmentation, other explanations could be a lack of strategic development, resulting in difficulties to bring research into clinical settings. Providing well-accepted user interaction is often more challenging to solve than the AI algorithm powering the DDSS's cognition. Nevertheless, the evaluation dimension indicates that most papers focus on evaluating the cognition by exclusively using accuracy scores. The human interaction with those systems is often neglected. This demonstrates a significant research gap when it comes to the enrichment of clinical processes with IT. Research on AI and DDSSs should perhaps focus more on the effects of the clinical processes, similar to health technologies, that are more successful in clinical settings, e.g. diagnostic imaging.

4.2.2. Trustworthiness of AI and Digital Decision Support Systems in Psychiatry – RQ2

We answer RQ2 "trustworthiness" by investigating evaluation methods, accuracy, possible reasons for biases, and other ways which could lead to wrong recommendations of AI and DDSSs in psychiatry. We found that the majority of articles investigated neither statistical significance (present in 31/80 papers) nor possible reasons for biases (27/80). Ranking according to maturity levels revealed that research mainly dealt with algorithm development. In contrast, randomized control trials were rare, and only two systems in production were found. The fact that neither the concepts of decision support systems (Power, 2008, pp. 121–140) nor AI (Yu et al., 2018) are new indicates that the field is both stagnating and getting increasingly complex. This fact can be observed in current research and is also present in commercial products like IBM Watson (Strickland, 2019). One crucial factor that has held back AI in the past has recently been overcome: the lack of computing power. This positive development has led to many AI-related success stories in areas like finance, retail, or marketing. Notable examples include Google or Amazon, which are heavily dependent on AI (Smith & Linden, 2017). However, the medical domain is complex, the generalizability of diagnoses is questionable, while data collection and reuse are time-consuming and expensive. Additionally, many requirements concerning data protection and anonymization pose difficulties. Also, diagnosing patients is not a straightforward matter, especially in mental health, and the reproducibility of a patient diagnosis by humans is low (Aboraya et al., 2006; Basco et al., 2000; Mendel et al., 2011; Muller, 2013). Studies suggest that not only diagnostic, but also administrative errors, run rampant in modern-day diagnoses (Davis et al., 2016). This means that AI's training or labeling data itself is probably inconsistent, making establishing a ground truth for ML training difficult. In this context, computerized systems function as catalysts for already present errors in the dataset by enabling the large-scale reproduction of already biased decisions. It is questionable whether simple algorithms like SVMs (used in 24/80 papers) or logistic

regression (19/80) can model complex neurophysiology processes present in the human brain to a satisfactory degree.

Evaluation is an important keyword when it comes to the question of trustworthiness. Data scientists and AI researchers focus on improving accuracy scores since the academic community has decided that this constitutes the primary criterion for success. Other evaluation metrics are often neglected. According to our review, 34 out of the 80 papers focus on accuracy evaluations. A mean accuracy of 82.8% seems high. However, accuracy alone is insufficient to measure whether AI algorithms or DDSSs perform well. One crucial shortcoming of accuracy scores is that they are highly dependent on both sampling and the number of positive and negative cases that are present in the evaluation sets. In contrast, the class balance must be preserved to get realistic accuracy scores. By default, classification problems concerning psychiatric diseases are highly unbalanced, meaning the number of negative cases in a random population sample is much higher than the number of positive ones. Additionally, since researchers make evaluations based on their personally collected data, they may reproduce their own biases. Due to the problems mentioned above, the resulting accuracy values make comparisons between different research results difficult. Benchmarking different approaches becomes even more complicated when only a single measure like accuracy or AUC values are mentioned, as it was the case in most of the investigated research.

4.2.3. Adoption of New Technology - RQ3

When looking at RQ3 "adoption of new technology", our literature review uncovered that DDSS' input data is mostly based on checklists or questionnaires. Checklists and questionnaires as a tool for diagnosis have been tested and validated over the past decades and are still the only mentioned approach for diagnosis in medical guidelines like NICE (NICE Guideline NG116, 2020). New data sources that could be used for digital phenotyping have still not been broadly adopted. Wider use of internationally recognized taxonomies for clinical notes and standardization of data capture would also improve the performance and quality of AI and DDSSs. Since new technology has a short lifecycle, it is more difficult to find evidence supporting their clinical use, which might explain why there is less research. From an algorithmic perspective, the most common algorithm used in articles of this review was SVM. Compared to other algorithms, SVMs have less stringent assumptions for input data and are easy to implement. However, SVMs are not explainable by default. Explainable AI (XAI) is still a neglected topic of the current state of the art in health informatics. From a legal (article 13 and 14 of the EU General Data Protection Regulation) as well as from an ethical point of view, it is essential to understand why systems produce certain outputs (Safdar et al., 2020). Since cases are known where the application of AI algorithms has resulted in discrimination based on ethnicity or gender (Buolamwini & Gebru, 2018; Leavy, 2018), decision transparency is very much needed, especially in a sensitive domain like healthcare. XAI has the potential to bring accountability, transparency, and traceable results for DDSSs (Pawar et al., 2020) by enabling contestability of AI-based decisions (Ploug & Holm, 2020). A detailed definition of XAI, opportunities, and challenges are described by Barredo Arrieta et al. (2020).

The detection of causality also plays a major role in DDSS success. Currently used DDSS algorithms do not interpret causes and effects, thus are not able to detect why certain associations and correlations exist. This limits AI in being able to generalize beyond its narrow domains and transfer its skills to different problems.

4.2.4. Ways to Improve AI and Digital Decision Support Systems in Psychiatry – RQ4

Advancements of AI algorithms and DDSSs are highly dependent on data availability. As shown above, data impacts AI models' training and is also the primary source of evaluation and benchmarking. We think that many current problems are unrelated to algorithms' cognition or intelligence but can be better explained by a lack of high-quality data.

We propose that further research dedicates renewed attention to the use of unobtrusive data by DDSSs to supplement diagnostic data and clinical questionnaires. This shifts the focus from a diagnosis' perspective based on generalized artificial categorization back to physiological problems caused by different diseases giving better insights into the possible cause of mental illness and effective therapeutic intervention. Unobtrusive data is not impacted by current issues in healthcare concerning data and diagnosis standardization and data collection. Additionally, it also helps to mitigate potential biases in available data sources like electronic health records or checklists/questionaries.

To ensure that AI and DDSSs are less biased in psychiatric research, we propose using a unified benchmark dataset. Such datasets should contain anonymized, open-access data from many different resources. A unified benchmark can help to overcome challenges in measuring the correctness of DDSS algorithms. It helps to obtain standardized benchmark results, which makes the comparison of different approaches possible. This is already common in other disciplines where AI is used; examples include the MNIST Database for handwritten digits by the US National Institute of Standards and Technology (LeCun et al., 1999) or TweetEval for Tweet classification (Barbieri et al., 2020). Further research needs to specify how such unified benchmark datasets for DDSS and AI in psychiatry could look like.

Apart from relying on more and better data, we want to highlight the importance of adding confusion matrices in academic papers. A confusion matrix helps to make the performance of different algorithms comparable. It is impossible to produce aggregated *meta*-analyses to evaluate the performance of DDSS and AI when accuracy, AUC values, or F1 values are the only measures that researchers calculate since these values cannot be converted to a unified metric. Nevertheless, we found that only 42 papers listed confusion matrixes.

Given that most studies in this review used already available datasets which were not sampled individually for their research (maturity level two or lower), their high accuracy values must not necessarily reflect on a good performance of level four or higher DDSSs. We suggest that DDSSs need to be tested in a clinical setting to evaluate their real-world performance and efficacy. Additionally, the high accuracy of the algorithms could be an indication of overfitting. Current research mostly neglects this.

Nevertheless, accurate predictions alone are not always sufficient for medical decision-making. We argue that one success factor of DDSSs in healthcare is understanding causality and dealing with counterfactuals. This argument is also supported by popular scientists like Judea Pearl (Pearl & Mackenzie, 2018). New approaches in the field of AI, like causal representation learning (Scholkopf et al., 2021), could help overcome these challenges.

It is not sufficient to focus exclusively on one dimension of our framework like data or decision technology. In order to introduce DDSSs and AI safely into clinical practice, there is a need for standardization and unified evaluation criteria for every part of the decision support system framework by Bertl et al. (2020). A standardized way of evaluating DDSSs in clinical practice is needed to show the unbiased performance of DDSSs and AI in healthcare. At the moment, this has been neglected by the scientific community. However, a standardized evaluation is the foundation of the trustworthiness of computerized decision-making.

4.2.5. Limitations

This survey has several limitations. First, we only included peer-reviewed publications in English. Relevant DDSSs might have been published as pre-prints or news reports. DDSSs may also have been implemented in real-world clinical practice without previously publishing these systems in academic journals. We think that this concern is unlikely but still possible. One example of this would be the product EBMeDS (Duodecim Medical Publications Ltd., 2020). Several publications about EBMeDS exist, but none of them mentions psychiatry, although it is used in this area in production. These points may partly

explain why we found low maturity scores.

Not the least, further research needs to be done to analyze maturity scores of currently used DDSSs. Since research about DDSSs in psychiatry is limited, our results are based on a small corpus of literature. This needs to be taken into account while interpreting the quantitative results of this survey.

Regarding our research question on trustworthiness, it needs to be noted that this is a highly subjective matter. Indeed, no unified measurement has evolved. We try to quantify trustworthiness by looking at maturity and accuracy scores. Further research could use other approaches like surveys of stakeholders to measure their opinion.

Another limitation is that we calculate statistical parameters over the whole study population. We argue that this is valid because all our papers still belong to the top-level category of DDSSs in psychiatry, although they might follow different approaches and deal with various diseases. The overview provided by this survey is needed to establish a baseline on how well DDSSs in psychiatry work generally. Especially the aggregation of sample size used for training and evaluation, as well as the accuracy scores can be an indicator for trustworthiness and is therefore highly relevant for our research. We accept that our aggregations might hide important variations in the data. A more granular sub-group analysis would be desirable but is unlikely to find statistically significant results due to the even smaller corpus of the literature. Here, we encourage further research. Nevertheless, such studies still cannot conclude on the general subject area of DDSSs in psychiatry.

5. Conclusion

There is no evidence of widespread usage of AI or DDSSs applications in psychiatry or other clinical specialties in everyday practice. Although the algorithms' high accuracy scores seem to support their use, this systematic literature review indicated problems concerning small sample sizes, possibilities of bias, lack of evaluation in production, and potential difficulties in establishing a ground truth. One reason that could explain why collecting new and original data is difficult might be the absence of standardization, centralized eHealth infrastructure, and data reuse capabilities in healthcare systems. Concepts to cope with health data fragmentation are needed as well as concepts to ensure data quality. Additionally, we are missing broad evidence of AI's and DDSSs' success confirmed by clinical studies to justify large-scale adoption. We also advocate for introducing a standardized concept for evaluating DDSS and AI in healthcare. One such component could include carefully assembled unified benchmark datasets to establish a consistent way of evaluating algorithmic accuracy of DDSSs and AI algorithms and helping to reduce bias. Algorithmic bias reflects human biases and culture, possibly with catastrophic consequences. On the other hand, a well-consolidated AI system that discovers causations in big data could amass the kind of knowledge of human mental disorders, their genetic origins and expressions in diagnoses and human behavior in a way that no other method can – potentially tapping into the 'source code of the mind' on the deepest neurocognitive levels. However, although much needed, we see the opportunity for AI and DDSSs to improve psychiatry at the moment to remain just that – an opportunity for the future.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Markus Bertl: Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – original draft. **Peeter Ross:** Conceptualization, Resources, Writing – review & editing, Supervision. **Dirk Draheim:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

The following two sections list the literature used for this review:

Publication Title (Psychiatry)	Year
A clinical risk stratification tool for predicting treatment resistance in major depressive disorder (Perlis, 2013)	2013
A deformable registration method for automated morphometry of MRI brain images in neuropsychiatric research (Schwarz et al., 2007)	2007
A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy (Khodayari-Rostamabad et al., 2010)	2010
A risk calculator to predict adult attention-deficit/hyperactivity disorder: Generation and external validation in three birth cohorts and one clinical sample (Caye et al., 2019)	2019
A situation-aware system for the detection of motion disorders of patients with Autism Spectrum Disorders (Coronato et al., 2014)	2014
A web-based clinical decision tool to support treatment decision-making in psychiatry: A pilot focus group study with clinicians, patients and carers (Henshall et al., 2017)	2017
Automatic recognition of symptom severity from psychiatric evaluation records (Goodwin et al., 2017)	2017
Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: The Smartcrises study protocol (Berrouiguet et al., 2019)	2019
Computational neuroimaging strategies for single patient predictions (Stephan et al., 2017)	2017
Computer-aided DSM-IV-diagnostics - Acceptance, use and perceived usefulness in relation to users' learning styles (Bergman & Fors, 2005)	2005
Design and methods of the 'monitoring outcomes of psychiatric pharmacotherapy' (MOPHAR) monitoring program - A study protocol (Simoons et al., 2019)	2019
Discrimination of schizophrenia auditory hallucinators by machine learning of resting-state functional MRI (Chyzhyk et al., 2015)	2015
Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data (Zhao & So, 2019)	2019
Drug side effect extraction from clinical narratives of psychiatry and psychology patients (Sohn et al., 2011)	2011
From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support (Constantinou et al., 2016)	2016

(*continued on next page*)

(*continued*)

Publication Title (Psychiatry)	Year
Implementing a digital clinical decision support tool for side effects of antipsychotics: A focus group study (Henshall et al., 2019)	2019
Leveraging the utility of pharmacogenomics in psychiatry through clinical decision support: A focus group study (Goodspeed et al., 2019)	2019
Machine learning methods to predict child posttraumatic stress: A proof of concept study (Saxe et al., 2017)	2017
Multimodel decision support system for psychiatry problem (Suhasini et al., 2011)	2011
Predicting patient outcomes in psychiatric hospitals with routine data: A machine learning approach (Wolff et al., 2020)	2020
Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare (Hatton et al., 2019)	2019
Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records (Posada et al., 2017)	2017
The development and evaluation of a computerized decision aid for the treatment of psychotic disorders (Tasma et al., 2018)	2018
The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes (Jones & Bennell, 2007)	2007
Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition (Koutsouleris et al., 2009)	2009
VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry (Mane et al., 2012)	2012
A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data (Ambrosen et al., 2020)	2020
A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder (Perez Arribas et al., 2018)	2018
An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: Application for schizophrenia detection (Kacur et al., 2020)	2020
An Ensemble Approach to Predict Schizophrenia Using Protein Data in the N-methyl-D-Aspartate Receptor (NMDAR) and Tryptophan Catabolic Pathways (Lin et al., 2020)	2020
Automated classification of fMRI during cognitive control identifies more severely disorganized subjects with schizophrenia (Yoon et al., 2012)	2012
Automated Depression Detection Using Deep Representation and Sequence Learning with EEG Signals (Ay et al., 2019)	2019
Counting trees in Random Forests: Predicting symptom severity in psychiatric intake reports (Scheurwegs et al., 2017)	2017
Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder (Choi et al., 2020)	2020
Delirium misdiagnosis risk in psychiatry: A machine learning-logistic regression predictive algorithm (Hercus & Hudaib, 2020)	2020
Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms (Sun et al., 2009)	2009
EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency (G. R. Alam et al., 2016)	2016
Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization (Papini et al., 2018)	2018
Estimation of the Development of Depression and PTSD in Children Exposed to Sexual Abuse and Development of Decision Support Systems by Using Artificial Intelligence (Ucuz et al., 2020)	2020
Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning (Tazawa et al., 2020)	2020
Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services (Røst et al., 2020)	2020
Machine-based classification of ADHD and nonADHD participants using time/frequency features of event-related neuroelectric activity (Öztoprak et al., 2017)	2017
Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder (Frick et al., 2020)	2020
Predicting individual clinical trajectories of depression with generative embedding (Frässle et al., 2020)	2020
Psychiatric comorbid disorders of cognition: A machine learning approach using 1175 UK Biobank participants (Li et al., 2020)	2020
Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: Feasibility study (Haines-Delmont et al., 2020)	2020
Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction (Benrimoh et al., 2020)	2020
Using structural MRI to identify individuals at genetic risk for bipolar disorders: A 2-cohort, machine learning study (Hajek et al., 2015)	2015
Vocal pattern detection of depression among older adults (M. Smith et al., 2020)	2020
Web of objects based ambient assisted living framework for emergency psychiatric state prediction (M. G. R. Alam et al., 2016)	2016

Publication Title (PTSD)	Year
A First Step towards a Clinical Decision Support System for Post-traumatic Stress Disorders (Ma et al., 2016)	2016
A mobile app for patients and those who care about them: A case study for veterans with PTSD + anger (Barish et al., 2019)	2019
A multimodal approach for predicting changes in PTSD symptom severity (Mallol-Ragolta et al., 2018)	2018
A neural network based model for predicting psychological conditions (Dabek & Caban, 2015a)	2015
A wearable health monitoring system for posttraumatic stress disorder (McWhorter et al., 2017)	2017
An alternative evaluation of post traumatic stress disorder with machine learning methods (Omurca & Ekinci, 2015)	2015
Bridging a translational gap: Using machine learning to improve the prediction of PTSD (Karstoft, Galatzer-Levy, et al., 2015)	2015
Investigating voice quality as a speaker-independent indicator of depression and PTSD (Scherer et al., 2013)	2013
Machine learning methods to predict child posttraumatic stress: A proof of concept study (Saxe et al., 2017)	2016
Measuring post traumatic stress disorder in twitter (Coppersmith et al., 2014)	2014

(*continued on next page*)

(*continued*)

Publication Title (PTSD)	Year
Quantitative forecasting of PTSD from early trauma responses: A Machine Learning application (Galatzer-Levy et al., 2014)	2014
Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews (Scherer et al., 2016)	2016
Technology-Enhanced Stepped Collaborative Care Targeting Posttraumatic Stress Disorder and Comorbidity After Injury: A Randomized Controlled Trial (Zatzick et al., 2015)	2015
Towards clinical decision support for veteran mental health crisis events using tree algorithm (Hossain et al., 2019)	2019
A voice-based automated system for PTSD screening and monitoring (Xu et al., 2012)	2012
Automated Assessment of Patients' Self-Narratives for Posttraumatic Stress Disorder Screening Using Natural Language Processing and Text Mining (He et al., 2017)	2017
Beyond symptom self-report: use of a computer "avatar" to assess post-traumatic stress disorder (PTSD) symptoms (Myers et al., 2016)	2016
Customized computer-based administration of the PCL-5 for the efficient assessment of PTSD: A proof-of-principle study (Finkelman et al., 2017)	2017
Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers (Karstoft, Statnikov, et al., 2015)	2015
Feasibility, acceptability, and potential efficacy of the PTSD Coach app: A pilot randomized controlled trial with community trauma survivors (Miner et al., 2016)	2016
Heart rate variability: Pre-deployment predictor of post-deployment PTSD symptoms (Pyne et al., 2016)	2016
Improving speech-based PTSD detection via multi-view learning (Zhuang et al., 2014)	2014
Leveraging Big Data to Model the Likelihood of Developing Psychological Conditions After a Concussion (Dabek & Caban, 2015b)	2015
Linguistic predictors of trauma pathology and physical health (Alvarez-Conrad et al., 2001)	2001
Physiology-Driven Adaptive Virtual Reality Stimulation for Prevention and Treatment of Stress Related Disorders (Ćosić et al., 2010)	2010
Posttraumatic Stress Disorder: Diagnostic Data Analysis by Data Mining Methodology (Marinić et al., 2007)	2007
Preliminary Evaluation of PTSD Coach, a Smartphone App for Post-Traumatic Stress Symptoms (Kuhn et al., 2014)	2014
Temporal analysis of heart rate variability as a predictor of post traumatic stress disorder in road traffic accidents survivors (Shaikh al arab et al., 2012)	2012
The use of immersive virtual reality (VR) to predict the occurrence 6 months later of paranoid thinking and posttraumatic stress symptoms assessed by self-report and interviewer methods: A study of individuals who have been physically assaulted. (Freeman et al., 2014)	2014
Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level (Gong et al., 2014)	2014

References

Aboraya, A., Rankin, E., France, C., El-Missiry, A., & John, C. (2006). The Reliability of Psychiatric Diagnosis Revisited. *Psychiatry (Edgmont), 3*(1), 41–50.

Alam, G. R., Haw, R., Kim, S. S., Azad, A. K., Abedin, S. F., & Hong, C. S. (2016). EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency. *IEEE Transactions on Industrial Informatics*, *12*(6), 2321–2330. Scopus. 10.1109/TII.2016.2610191.

Alam, M. G. R., Abedin, S. F., Ameen, M. A., & Hong, C. S. (2016). Web of objects based ambient assisted living framework for emergency psychiatric state prediction. *Sensors (Switzerland), 16*(9), Scopus. https://doi.org/10.3390/s16091431

Al-Huthail, Y. R. (2008). Accuracy of Referring Psychiatric Diagnosis. *International Journal of Health Sciences, 2*(1), 35–38.

AlSalem, M., AlHarbi, M. A., Badeghiesh, A., & Tourian, L. (2020). Accuracy of initial psychiatric diagnoses given by nonpsychiatric physicians: A retrospective chart review. *Medicine, 99*(51), Article e23708. https://doi.org/10.1097/MD.0000000000023708

Alvarez-Conrad, J., Zoellner, L. A., & Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology, 15*(7), S159–S170. https://doi.org/10.1002/acp.839

Ambrosen, K. S., Skjerbæk, M. W., Foldager, J., Axelsen, M. C., Bak, N., Arvastson, L., Christensen, S. R., Johansen, L. B., Raghava, J. M., Oranje, B., Rostrup, E., Nielsen, M. Ø., Osler, M., Fagerlund, B., Pantelis, C., Kinon, B. J., Glenthøj, B. Y., Hansen, L. K., & Ebdrup, B. H. (2020). A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Translational Psychiatry*, *10*(1). Scopus. 10.1038/s41398-020-00962-8.

Ay, B., Yildirim, O., Talo, M., Baloglu, U. B., Aydin, G., Puthankattil, S. D., & Acharya, U. R. (2019). Automated Depression Detection Using Deep Representation and Sequence Learning with EEG Signals. *Journal of Medical Systems, 43*(7), Scopus. https://doi.org/10.1007/s10916-019-1345-y

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. ArXiv: 2010.12421 [Cs] http://arxiv.org/abs/2010.12421.

Barish, G., Aralis, H., Elbogen, E., & Lester, P. (2019). A mobile app for patients and those who care about them: A case study for veterans with PTSD + anger. *ACM Int. Conf. Proc. Ser.*, 1–10. Scopus. 10.1145/3329189.3329248.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Basco, M. R., Bostic, J. Q., Davies, D., Rush, A. J., Witte, B., Hendrickse, W., & Barnett, V. (2000). Methods to Improve Diagnostic Accuracy in a Community Mental Health Setting. *American Journal of Psychiatry, 157*(10), 1599–1605. https://doi.org/10.1176/appi.ajp.157.10.1599

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., Spurr, C., Khorasani, R., Tanasijevic, M., & Middleton, B. (2003). 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. https://doi.org/10.1197/jamia.M1370

Benrimoh, D., Tanguay-Sela, M., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., Fratila, R., Parikh, S. V., Karp, J. F., Heller, K., Vahia, I. V., Blumberger, D. M., Karama, S., Vigod, S. N., Myhr, G., Martins, R., Rollins, C., Popescu, C., Lundrigan, E., … Margolese, H. C. (2020). Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction. *BJPsych Open, 7*(1). Scopus. 10.1192/bjo.2020.127.

Bergman, L. G., & Fors, U. G. H. (2005). Computer-aided DSM-IV-diagnostics—Acceptance, use and perceived usefulness in relation to users' learning styles. *BMC Medical Informatics and Decision Making, 5*, Scopus. https://doi.org/10.1186/1472-6947-5-1

Berrouiguet, S., Barrigón, M. L., Castroman, J. L., Courtet, P., Artés-Rodríguez, A., & Baca-García, E. (2019). Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: The Smartcrises study protocol. *BMC Psychiatry, 19*(1), Scopus. https://doi.org/10.1186/s12888-019-2260-y

Bertl, M. (2019). News analysis for the detection of cyber security issues in digital healthcare: *Young. Information Scientist, 4*, 1–15. https://doi.org/10.25365/yis-2019-4-1

Bertl, M., Metsallik, J., & Ross, P. (2020). *Digital Decision Support Systems for Post-Traumatic Stress Disorder—Implementing a novel framework for decision support systems based on a technology-focused, systematic literature review*. https://doi.org/10.13140/RG.2.2.12571.28965/1.

Bettinger, C. J. (2018). Advances in Materials and Structures for Ingestible Electromechanical Medical Devices. *Angewandte Chemie International Edition, 57*(52), 16946–16958. https://doi.org/10.1002/anie.201806470

Boza, A., Ortiz, A., Vicens, E., & Poler, R. (2009). A Framework for a Decision Support System in a Hierarchical Extended Enterprise Decision Context. In R. Poler, M. van Sinderen, & R. Sanchis (Eds.), *Enterprise Interoperability* (pp. 113–124). Springer. 10.1007/978-3-642-04750-3_10.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html.

Camacho, J., Zanoletti-Mannello, M., Landis-Lewis, Z., Kane-Gill, S. L., & Boyce, R. D. (2020). A Conceptual Framework to Study the Implementation of Clinical Decision Support Systems (BEAR): Literature Review and Concept Mapping. *Journal of Medical Internet Research, 22*(8), Article e18388. https://doi.org/10.2196/18388

Caye, A., Agnew-Blais, J., Arseneault, L., Gonçalves, H., Kieling, C., Langley, K., Menezes, A. M. B., Moffitt, T. E., Passos, I. C., Rocha, T. B., Sibley, M. H., Swanson, J. M., Thapar, A., Wehrmeister, F., & Rohde, L. A. (2019). A risk calculator to predict adult attention-deficit/hyperactivity disorder: Generation and external validation in three birth cohorts and one clinical sample. *Epidemiology and Psychiatric Sciences*. Scopus. 10.1017/S2045796019000283.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety, 28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Choi, B., Shim, G., Jeong, B., & Jo, S. (2020). Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Scientific Reports, 10*(1), Scopus. https://doi.org/10.1038/s41598-020-64709-7

Chyzhyk, D., Graña, M., Öngür, D., & Shinn, A. K. (2015). Discrimination of schizophrenia auditory hallucinators by machine learning of resting-state functional MRI. *International Journal of Neural Systems, 25*(3), Scopus. https://doi.org/10.1142/S0129065715500070

Constantinou, A. C., Fenton, N., Marsh, W., & Radlinski, L. (2016). From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support. *Artificial Intelligence in Medicine, 67*, 75–93. Scopus. 10.1016/j.artmed.2016.01.002.

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. *Proc. Int. Conf. Weblogs Soc. Media, ICWSM, 579–582*, Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84909982111&partnerID=40&md5=385e0d4e8cb2b52e387d55317b1ac262.

Coronato, A., De Pietro, G., & Paragliola, G. (2014). A situation-aware system for the detection of motion disorders of patients with Autism Spectrum Disorders. *Expert Systems with Applications, 41*(17), 7868–7877. Scopus. 10.1016/j.eswa.2014.05.011.

Ćosić, K., Popović, S., Kukolja, D., Horvat, M., & Dropuljić, B. (2010). Physiology-Driven Adaptive Virtual Reality Stimulation for Prevention and Treatment of Stress Related Disorders. *Cyberpsychology, Behavior, and Social Networking, 13*(1), 73–78. https://doi.org/10.1089/cyber.2009.0260

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry, 13*(1), 28–35. https://doi.org/10.1002/wps.20087

Dabek, F., & Caban, J. J. (2015a). *A neural network based model for predicting psychological conditions* (Vol. 9250). https://doi.org/10.1007/978-3-319-23344-4_25

Dabek, F., & Caban, J. J. (2015b). Leveraging big data to model the likelihood of developing psychological conditions after a concussion. In Roy A., Venayagamoorthy K., Alimi A., Angelov P., & Trafalis T. (Eds.), *Procedia Comput. Sci.* (Vol. 53, pp. 265–273). Elsevier B.V.; Scopus. 10.1016/j.procs.2015.07.303.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data, 6*(1), 54. https://doi.org/10.1186/s40537-019-0217-0

Davis, K. A. S., Sudlow, C. L. M., & Hotopf, M. (2016). Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry, 16*(1), 263. https://doi.org/10.1186/s12888-016-0963-x

Day, S., Seninger, C., Fan, J., Pundi, K., Perino, A., & Turakhia, M. (2019). *Digital Health Consumer Adoption Report 2019*. Stanford Medicine.

Duodecim Medical Publications Ltd. (2020). *EBMEDS White Paper*. https://www.ebmeds.org/wp-content/uploads/sites/16/2020/10/WhitePaper_2020-1.pdf.

eHealth, Well-being, and Ageing (Unit H.3). (2019). *EHealth adoption in primary healthcare in the EU is on the rise* [Text]. European Commission. https://ec.europa.eu/digital-single-market/en/news/ehealth-adoption-primary-healthcare-eu-rise.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Finkelman, M. D., Lowe, S. R., Kim, W., Gruebner, O., Smits, N., & Galea, S. (2017). Customized computer-based administration of the PCL-5 for the efficient assessment of PTSD: A proof-of-principle study. *Psychological Trauma: Theory, Research, Practice, and Policy, 9*(3), 379–389. https://doi.org/10.1037/tra0000226

Frässle, S., Marquand, A. F., Schmaal, L., Dinga, R., Veltman, D. J., van der Wee, N. J. A., van Tol, M.-J., Schöbi, D., Penninx, B. W. J. H., & Stephan, K. E. (2020). Predicting individual clinical trajectories of depression with generative embedding. *NeuroImage: Clinical, 26*, Scopus. https://doi.org/10.1016/j.nicl.2020.102213

Freeman, D., Antley, A., Ehlers, A., Dunn, G., Thompson, C., Vorontsova, N., Garety, P., Kuipers, E., Glucksman, E., & Slater, M. (2014). The use of immersive virtual reality (VR) to predict the occurrence 6 months later of paranoid thinking and posttraumatic stress symptoms assessed by self-report and interviewer methods: A study of individuals who have been physically assaulted. *Psychological Assessment, 26*(3), 841–847. https://doi.org/10.1037/a0036240

Frick, A., Engman, J., Alaie, I., Björkstrand, J., Gingnell, M., Larsson, E.-M., Eriksson, E., Wahlstedt, K., Fredrikson, M., & Furmark, T. (2020). Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. *Journal of Affective Disorders, 261*, 230–237. Scopus. 10.1016/j.jad.2019.10.027.

Galatzer-Levy, I. R., Karstoft, K.-I., Statnikov, A., & Shalev, A. Y. (2014). Quantitative forecasting of PTSD from early trauma responses: A Machine Learning application. *Journal of Psychiatric Research, 59*, 68–76. Scopus. 10.1016/j.jpsychires.2014.08.017.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *Npj Digital Medicine, 4*(1), 1–8. https://doi.org/10.1038/s41746-021-00385-9

Gong, Q., Li, L., Tognin, S., Wu, Q., Pettersson-Yeo, W., Lui, S., Huang, X., Marquand, A. F., & Mechelli, A. (2014). Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level. *Psychological Medicine, 44*(1), 195–203. https://doi.org/10.1017/S0033291713000561

Goodspeed, A., Kostman, N., Kriete, T. E., Longtine, J. W., Smith, S. M., Marshall, P., Williams, W., Clark, C., & Blakeslee, W. W. (2019). Leveraging the utility of pharmacogenomics in psychiatry through clinical decision support: A focus group study. *Annals of General Psychiatry, 18*(1), Scopus. https://doi.org/10.1186/s12991-019-0237-3

Goodwin, T. R., Maldonado, R., & Harabagiu, S. M. (2017). Automatic recognition of symptom severity from psychiatric evaluation records. *Journal of Biomedical Informatics, 75*, S71–S84. Scopus. 10.1016/j.jbi.2017.05.020.

Greenes, R. A., Bates, D. W., Kawamoto, K., Middleton, B., Osheroff, J., & Shahar, Y. (2018). Clinical decision support models and frameworks: Seeking to address research issues underlying implementation successes and failures. *Journal of Biomedical Informatics, 78*, 134–143. https://doi.org/10.1016/j.jbi.2017.12.005

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., … Olesen, J. (2011). Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology, 21*(10), 718–779. https://doi.org/10.1016/j.euroneuro.2011.08.008

Haines-Delmont, A., Chahal, G., Bruen, A. J., Wall, A., Khan, C. T., Sadashiv, R., & Fearnley, D. (2020). Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: Feasibility study. *JMIR MHealth and UHealth, 8*(6), Scopus. https://doi.org/10.2196/15901

Hajek, T., Cooke, C., Kopecek, M., Novak, T., Hoschl, C., & Alda, M. (2015). Using structural MRI to identify individuals at genetic risk for bipolar disorders: A 2-cohort, machine learning study. *Journal of Psychiatry and Neuroscience, 40*(5), 316–324. Scopus. 10.1503/jpn.140142.

Hamidia, A., Kheirkhah, F., Chehrazi, M., Basirat, Z., Ghadimi, R., Barat, S., Cuijpers, P., O'Connor, E., Mirtabar, S. M., & Faramarzi, M. (2022). Screening of psychiatric disorders in women with high-risk pregnancy: Accuracy of three psychological tools. *Health Science Reports, 5*(2), Article e518. https://doi.org/10.1002/hsr2.518

Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders, 246*, 857–860. Scopus. 10.1016/j.jad.2018.12.095.

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated Assessment of Patients' Self-Narratives for Posttraumatic Stress Disorder Screening Using Natural Language Processing and Text Mining. *Assessment, 24*(2), 157–172. https://doi.org/10.1177/1073191115602551

Henshall, C., Cipriani, A., Ruvolo, D., Macdonald, O., Wolters, L., & Koychev, I. (2019). Implementing a digital clinical decision support tool for side effects of antipsychotics: A focus group study. *Evidence-Based Mental Health, 22*(2), 56–60. Scopus. 10.1136/ebmental-2019-300086.

Henshall, C., Marzano, L., Smith, K., Attenburrow, M.-J., Puntis, S., Zlodre, J., Kelly, K., Broome, M. R., Shaw, S., Barrera, A., Molodynski, A., Reid, A., Geddes, J. R., & Cipriani, A. (2017). A web-based clinical decision tool to support treatment decision-making in psychiatry: A pilot focus group study with clinicians, patients and carers. *BMC Psychiatry, 17*(1), Scopus. https://doi.org/10.1186/s12888-017-1406-z

Hercus, C., & Hudaib, A.-R. (2020). Delirium misdiagnosis risk in psychiatry: A machine learning-logistic regression predictive algorithm. *BMC Health Services Research, 20*(1), Scopus. https://doi.org/10.1186/s12913-020-5005-1

Hossain, M. F., George, O., Johnson, N., Madiraju, P., Flower, M., Franco, Z., Hooyer, K., Rein, L., Mazaba, J. L., & Ahamed, S. I. (2019). Towards clinical decision support for veteran mental health crisis events using tree algorithm. In V. Getov, J.-.-L. Gaudiot, N. Yamai, S. Cimato, M. Chang, Y. Teranishi, J.-.-J. Yang, H. V. Leong, H. Shahriar, M. Takemoto, D. Towey, H. Takakura, A. Elci, S. Takeuchi, & S. Puri (Eds.), *Proc Int Comput Software Appl Conf* (Vol. 2, pp. 386–390). Scopus: IEEE Computer Society. https://doi.org/10.1109/COMPSAC.2019.10237.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jones, N. J., & Bennell, C. (2007). The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research, 11*(2), 219–233. Scopus. 10.1080/13811110701250176.

Kacur, J., Polec, J., Smolejova, E., & Heretik, A. (2020). An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: Application for schizophrenia detection. *IEEE Journal of Biomedical and Health Informatics, 24*(11), 3055–3065. Scopus. 10.1109/JBHI.2020.3002097.

Karstoft, K.-I., Galatzer-Levy, I. R., Statnikov, A., Li, Z., Shalev, A. Y., Ankri, Y., Freedman, S., Addesky, R., Israeli-Shalev, Y., Gilad, M., Roitman, P., & and For members of the Jerusalem Trauma Outreach and Prevention Study (J-TOPS) group. (2015). Bridging a translational gap: Using machine learning to improve the prediction of PTSD. *BMC Psychiatry, 15*(1). Scopus. 10.1186/s12888-015-0399-8.

Karstoft, K.-I., Statnikov, A., Andersen, S. B., Madsen, T., & Galatzer-Levy, I. R. (2015). Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers. *Journal of Affective Disorders, 184*, 170–175. https://doi.org/10.1016/j.jad.2015.05.057

Khodayari-Rostamabad, A., Hasey, G. M., MacCrimmon, D. J., Reilly, J. P., & Bruin, H. D. (2010). A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clinical Neurophysiology, 121*(12), 1998–2006. Scopus. 10.1016/j.clinph.2010.05.009.

Kitamura, T., Shima, S., Sakio, E., & Kato, M. (1989). Psychiatric Diagnosis in Japan. 2. Reliability of Conventional Diagnosis and Discrepancies with Research Diagnostic Criteria Diagnosis. *Psychopathology, 22*(5), 250–259. https://doi.org/10.1159/000284605

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. https://www.elsevier.com/_data/promis_misc/525444systematicreviewsguide.pdf.

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., & Gaser, C. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry*, *66*(7), 700–712. Scopus. 10.1001/archgenpsychiatry.2009.62.

Kuhn, E., Greene, C., Hoffman, J., Nguyen, T., Wald, L., Schmidt, J., Ramsey, K. M., & Ruzek, J. (2014). Preliminary Evaluation of PTSD Coach, a Smartphone App for Post-Traumatic Stress Symptoms. *Military Medicine, 179*(1), 12–18. https://doi.org/10.7205/MILMED-D-13-00271

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)* (pp. 14–16).

LeCun, Y., Cortes, C., & Burges, C. J. C. (1999). *MNIST handwritten digit database*. http://yann.lecun.com/exdb/mnist/.

Li, C., Gheorghe, D. A., Gallacher, J. E., & Bauermeister, S. (2020). Psychiatric comorbid disorders of cognition: A machine learning approach using 1175 UK Biobank participants. *Evidence-Based Mental Health*. Scopus. 10.1136/ebmental-2020-300147.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ, 339*, Article b2700. https://doi.org/10.1136/bmj.b2700

Lin, E., Lin, C.-H., Hung, C.-C., & Lane, H.-Y. (2020). An Ensemble Approach to Predict Schizophrenia Using Protein Data in the N-methyl-D-Aspartate Receptor (NMDAR). *and Tryptophan Catabolic Pathways. Frontiers in Bioengineering and Biotechnology, 8*, Scopus. https://doi.org/10.3389/fbioe.2020.00569

Ma, S., Galatzer-Levy, I. R., Wang, X., Fenyö, D., & Shalev, A. Y. (2016). A First Step towards a Clinical Decision Support System for Post-traumatic Stress Disorders. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2016*, 837–843. Scopus.

Mallol-Ragolta, A., Dhamija, S., & Boult, T. E. (2018). A multimodal approach for predicting changes in PTSD symptom severity. *ICMI - Proc. Int. Conf. Multimodal Interact., 324–333*, Scopus. https://doi.org/10.1145/3242969.3242981

Mane, K. K., Bizon, C., Schmitt, C., Owen, P., Burchett, B., Pietrobon, R., & Gersing, K. (2012). VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *Journal of Biomedical Informatics*, *45*(1), 101–106. Scopus. 10.1016/j.jbi.2011.09.003.

Marinić, I., Supek, F., Kovačić, Z., Rukavina, L., Jendričko, T., & Kozarić-Kovačić, D. (2007). Posttraumatic Stress Disorder: Diagnostic Data Analysis by Data Mining Methodology. *Croat Med J, 13*.

Maron, E., Baldwin, D. S., Balõtšev, R., Fabbri, C., Gaur, V., Hidalgo-Mazzei, D., … Eberhard, J. (2019). Manifesto for an international digital mental health network. *Digital Psychiatry, 2*(1), 14–24. https://doi.org/10.1080/2575517X.2019.1617575

McGlynn, E. A., Asch, S. M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., & Kerr, E. A. (2003). The Quality of Health Care Delivered to Adults in the United States. *New England Journal of Medicine, 348*(26), 2635–2645. https://doi.org/10.1056/NEJMsa022615

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U Test. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). American Cancer Society. 10.1002/9780470479216.corpsy0524.

McWhorter, J., Brown, L., & Khansa, L. (2017). A wearable health monitoring system for posttraumatic stress disorder. *Biologically Inspired Cognitive Architectures*, *22*, 44–50. Scopus. 10.1016/j.bica.2017.09.004.

Meltzer, E. C., Averbuch, T., Samet, J. H., Saitz, R., Jabbar, K., Lloyd-Travaglini, C., & Liebschutz, J. M. (2012). Discrepancy in diagnosis and treatment of post-traumatic stress disorder (PTSD): Treatment for the wrong reason. *The Journal of Behavioral Health Services & Research, 39*(2), 190–201. https://doi.org/10.1007/s11414-011-9263-x

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., Kissling, W., & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 9*.

Metsallik, J., Ross, P., Draheim, D., & Piho, G. (2018). Ten Years of the e-Health System in Estonia. *CEUR Workshop Proceedings, 10*.

Miner, A., Kuhn, E., Hoffman, J. E., Owen, J. E., Ruzek, J. I., & Taylor, C. B. (2016). Feasibility, acceptability, and potential efficacy of the PTSD Coach app: A pilot randomized controlled trial with community trauma survivors. *Psychological Trauma: Theory, Research, Practice, and Policy, 8*(3), 384–392. https://doi.org/10.1037/tra0000092

Mitchell, A. J., Vaze, A., & Rao, S. (2009). *Clinical diagnosis of depression in primary care: A meta-analysis*. Database of Abstracts of Reviews of Effects (DARE): Quality-Assessed Reviews. https://www.ncbi.nlm.nih.gov/books/NBK77945/.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Morselli, P. L., & Elgie, R. (2003). GAMIAN-Europe*/BEAM survey I – global analysis of a patient questionnaire circulated to 3450 members of 12 European advocacy groups operating in the field of mood disorders. *Bipolar Disorders, 5*(4), 265–278. https://doi.org/10.1034/j.1399-5618.2003.00037.x

Muller, R. J. (2013). *Doing Psychiatry Wrong: A Critical and Prescriptive Look at a Faltering Profession*. Routledge.

Myers, C. E., Radell, M. L., Shind, C., Ebanks-Williams, Y., Beck, K. D., & Gilbertson, M. W. (2016). Beyond symptom self-report: Use of a computer "avatar" to assess post-traumatic stress disorder (PTSD) symptoms. *Stress, 19*(6), 593–598. https://doi.org/10.1080/10253890.2016.1232385

Neuman, M. R., Baura, G. D., Meldrum, S., Soykan, O., Valentinuzzi, M. E., Leder, R. S., Micera, S., & Zhang, Y.-T. (2012). Advances in Medical Devices and Medical Electronics. *Proceedings of the IEEE*, *100*(Special Centennial Issue), 1537–1550. 10.1109/JPROC.2012.2190684.

Office-based Physician Electronic Health Record Adoption. (2019). Office of the National Coordinator for Health Information Technology. dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php.

Omurca, S. I., & Ekinci, E. (2015). An alternative evaluation of post traumatic stress disorder with machine learning methods. *INISTA - Int. Symp. Innov. Intell. SysT. Appl., Proc.* Scopus. 10.1109/INISTA.2015.7276754.

Öztoprak, H., Toycan, M., Alp, Y. K., Arıkan, O., Doğutepe, E., & Karakaş, S. (2017). Machine-based classification of ADHD and nonADHD participants using time/frequency features of event-related neuroelectric activity. *Clinical Neurophysiology, 128*(12), 2400–2410. Scopus. 10.1016/j.clinph.2017.09.105.

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *Npj Digital Medicine, 2*(1), 1–3. https://doi.org/10.1038/s41746-019-0155-4

Papini, S., Pisner, D., Shumake, J., Powers, M. B., Beevers, C. G., Rainey, E. E., Smits, J. A. J., & Warren, A. M. (2018). Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *Journal of Anxiety Disorders*, *60*, 35–42. Scopus. 10.1016/j.janxdis.2018.10.004.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). Data and its (dis)contents: A survey of dataset development and use in machine learning research. ArXiv:2012.05345 [Cs] http://arxiv.org/abs/2012.05345.

Pawar, U., O'Shea, D., Rea, S., & O'Reilly, R. (2020). Explainable AI in Healthcare. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 1–2. 10.1109/CyberSA49311.2020.9139655.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* ((1st ed.).). Basic Books.

Perez Arribas, I., Goodwin, G. M., Geddes, J. R., Lyons, T., & Saunders, K. E. A. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational. Psychiatry, 8*(1), Scopus. https://doi.org/10.1038/s41398-018-0334-0

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*(1), 7–14. Scopus. 10.1016/j.biopsych.2012.12.007.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics—A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine, 107*, Article 101901. https://doi.org/10.1016/j.artmed.2020.101901

Posada, J. D., Barda, A. J., Shi, L., Xue, D., Ruiz, V., Kuan, P.-H., Ryan, N. D., & Tsui, F. R. (2017). Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. *Journal of Biomedical Informatics*, *75*, S94–S104. Scopus. 10.1016/j.jbi.2017.05.019.

Post-traumatic stress disorder—[D] Evidence reviews for psychological, psychosocial and other non-pharmacological interventions for the treatment of PTSD in adults. (n.d.). National Institute for Health and Care Excellence. Retrieved September 9, 2020, from https://www.nice.org.uk/guidance/ng116/evidence/d-psychological-psychosocial-and-other-nonpharmacological-interventions-for-the-treatment-of-ptsd-in-adults-pdf-6602621008.

Power, D. J. (2008). Decision Support Systems: A Historical Overview. In F. Burstein & C. W. Holsapple (Eds.), *Handbook on Decision Support Systems 1: Basic Themes* (pp. 121–140). Springer. 10.1007/978-3-540-48713-5_7.

Pyne, J. M., Constans, J. I., Wiederhold, M. D., Gibson, D. P., Kimbrell, T., Kramer, T. L., Pitcock, J. A., Han, X., Williams, D. K., Chartrand, D., Gevirtz, R. N., Spira, J., Wiederhold, B. K., McCraty, R., & McCune, T. R. (2016). Heart rate variability: Pre-deployment predictor of post-deployment PTSD symptoms. *Biological Psychology, 121*, 91–98. https://doi.org/10.1016/j.biopsycho.2016.10.008

Røst, T. B., Clausen, C., Nytrø, O., Koposov, R., Leventhal, B., Westbye, O. S., Bakken, V., Flygel, L. H. K., Koochakpour, K., & Skokauskas, N. (2020). Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services. *Frontiers in Psychiatry*, *11*. Scopus. 10.3389/fpsyt.2020.564205.

Safdar, N. M., Banja, J. D., & Meltzer, C. C. (2020). Ethical considerations in artificial intelligence. *European Journal of Radiology, 122*, Article 108768. https://doi.org/10.1016/j.ejrad.2019.108768

Sauter, V. (1997). *Decision support systems: An applied managerial approach*. John Wiley & Sons, Inc.

Saxe, G. N., Ma, S., Ren, J., & Aliferis, C. (2017). Machine learning methods to predict child posttraumatic stress: A proof of concept study. *BMC Psychiatry, 17*(1), Scopus. https://doi.org/10.1186/s12888-017-1384-1

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A., & Morency, L.-P. (2016). Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews. *IEEE Transactions on Affective Computing, 7*(1), 59–73. Scopus. https://doi.org/10.1109/TAFFC.2015.2440264.

Scherer, S., Stratou, G., Gratch, J., & Morency, L.-P. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. *Proc. Annu. Conf. Int. Speech. Commun. Assoc., INTERSPEECH, 847–851*, Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84905232283&partnerID=40&md5=0511bf000f576d12ac1b8dbfcec3dac1.

Scheurwegs, E., Sushil, M., Tulkens, S., Daelemans, W., & Luyckx, K. (2017). Counting trees in Random Forests: Predicting symptom severity in psychiatric intake reports. *Journal of Biomedical Informatics*, *75*, S112–S119. Scopus. 10.1016/j.jbi.2017.06.007.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE, 109*(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

Schreyögg, J., Bäumler, M., & Busse, R. (2009). Balancing adoption and affordability of medical devices in Europe. *Health Policy, 92*(2), 218–224. https://doi.org/10.1016/j.healthpol.2009.03.016

Schwarz, D., Kasparek, T., Provaznik, I., & Jarkovsky, J. (2007). A deformable registration method for automated morphometry of MRI brain images in neuropsychiatric research. *IEEE Transactions on Medical Imaging*, *26*(4), 452–461. Scopus. 10.1109/TMI.2007.892512.

Shaikh al arab, A., Guédon-Moreau, L., Ducrocq, F., Molenda, S., Duhem, S., Salleron, J., Chaudieu, I., Bert, D., Libersa, C., & Vaiva, G. (2012). Temporal analysis of heart rate variability as a predictor of post traumatic stress disorder in road traffic accidents survivors. *Journal of Psychiatric Research*, *46*(6), 790–796. 10.1016/j.jpsychires.2012.02.006.

Sim, I., & Berlin, A. (2003). A Framework for Classifying Decision Support Systems. *AMIA Annual Symposium Proceedings, 2003*, 599–603.

Simoons, M., Ruhé, H. G., Van Roon, E. N., Schoevers, R. A., Bruggeman, R., Cath, D. C., Muis, D., Arends, J., Doornbos, B., & Mulder, H. (2019). Design and methods of the "monitoring outcomes of psychiatric pharmacotherapy" (MOPHAR) monitoring program—A study protocol. *BMC Health Services Research, 19*(1), Scopus. https://doi.org/10.1186/s12913-019-3951-2

Singh, T., & Rajput, M. (2006). *Misdiagnosis of Bipolar Disorder. Psychiatry (Edgmont), 3*(10), 57–63.

Sittig, D. F., Krall, M. A., Dykstra, R. H., Russell, A., & Chin, H. L. (2006). A survey of factors affecting clinician acceptance of clinical decision support. *BMC Medical Informatics and Decision Making, 6*(1), 6. https://doi.org/10.1186/1472-6947-6-6

Smith, B., & Linden, G. (2017). Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing, 21*(3), 12–18. https://doi.org/10.1109/MIC.2017.72

Smith, M., Dietrich, B. J., Bai, E.-W., & Bockholt, H. J. (2020). Vocal pattern detection of depression among older adults. *International Journal of Mental Health Nursing*, *29*(3), 440–449. Scopus. 10.1111/inm.12678.

Sohn, S., Kocher, J.-P. A., Chute, C. G., & Savova, G. K. (2011). Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, *18*(SUPPL. 1), 144–149. Scopus. 10.1136/amiajnl-2011-000351.

Spottswood, M., Davydow, D. S., & Huang, H. (2017). The Prevalence of Posttraumatic Stress Disorder in Primary Care: A Systematic Review. *Harvard Review of Psychiatry, 25*(4), 159–169. https://doi.org/10.1097/HRP.0000000000000136

Sprague, R. H. (1980). A Framework for the Development of Decision Support Systems. *MIS Quarterly, 4*(4), 1–26. https://doi.org/10.2307/248957

Stephan, K. E., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E. A., Brodersen, K. H., Rigoux, L., Moran, R. J., Daunizeau, J., Dolan, R. J., Friston, K. J., & Heinz, A. (2017). Computational neuroimaging strategies for single patient predictions. *NeuroImage*, *145*, 180–199. Scopus. 10.1016/j.neuroimage.2016.06.038.

Sterne, J. A. C., & Harbord, R. M. (2004). Funnel Plots in Meta-analysis. *The Stata Journal, 4*(2), 127–141. https://doi.org/10.1177/1536867X0400400204

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum, 56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Suhasini, A., Palanivel, S., & Ramalingam, V. (2011). Multimodel decision support system for psychiatry problem. *Expert Systems with Applications*, *38*(5), 4990–4997. Scopus. 10.1016/j.eswa.2010.09.152.

Sun, D., van Erp, T. G. M., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., Hardt, M. E., Nuechterlein, K. H., Toga, A. W., & Cannon, T. D. (2009). Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms. *Biological Psychiatry*, *66*(11), 1055–1060. Scopus. 10.1016/j.biopsych.2009.07.019.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *Npj Digital Medicine, 3*(1), 1–10. https://doi.org/10.1038/s41746-020-0221-y

Tasma, M., Roebroek, L. O., Lieburg, E. J., Knegtering, H., Delespaul, P. A., Boonstra, A., Swart, M., & Castelein, S. (2018). The development and evaluation of a computerized decision aid for the treatment of psychotic disorders. *BMC Psychiatry, 18*(1), Scopus. https://doi.org/10.1186/s12888-018-1750-7

Tazawa, Y., Liang, K.-C., Yoshimura, M., Kitazawa, M., Kaise, Y., Takamiya, A., Kishi, A., Horigome, T., Mitsukura, Y., Mimura, M., & Kishimoto, T. (2020). Evaluating depression with multimodal wristband-type wearable device: Screening and assessing patient severity utilizing machine-learning. *Heliyon, 6*(2), Scopus. https://doi.org/10.1016/j.heliyon.2020.e03274

Ucuz, I., Ari, A., Ozcan, O. O., Topaktas, O., Sarraf, M., & Dogan, O. (2020). Estimation of the Development of Depression and PTSD in Children Exposed to Sexual Abuse and Development of Decision Support Systems by Using Artificial Intelligence. *Journal of Child Sexual Abuse. Scopus.*. https://doi.org/10.1080/10538712.2020.1841350

Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., & Steinhausen, H.-C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology, 21*(9), 655–679. https://doi.org/10.1016/j.euroneuro.2011.07.018

Wolff, J., Gary, A., Jung, D., Normann, C., Kaier, K., Binder, H., Domschke, K., Klimke, A., & Franz, M. (2020). Predicting patient outcomes in psychiatric hospitals with routine data: A machine learning approach. *BMC Medical Informatics and Decision Making, 20*(1), Scopus. https://doi.org/10.1186/s12911-020-1042-2

Xu, R., Mei, G., Zhang, G., Gao, P., Judkins, T., Cannizzaro, M., & Li, J. (2012). A voice-based automated system for PTSD screening and monitoring. *Studies in Health Technology and Informatics, 173*, 552–558.

Yoon, J. H., Nguyen, D. V., McVay, L. M., Deramo, P., Minzenberg, M. J., Ragland, J. D., Niendham, T., Solomon, M., & Carter, C. S. (2012). Automated classification of fMRI during cognitive control identifies more severely disorganized subjects with schizophrenia. *Schizophrenia Research*, *135*(1–3), 28–33. Scopus. 10.1016/j.schres.2012.01.001.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zatzick, D., O'Connor, S. S., Russo, J., Wang, J., Bush, N., Love, J., Peterson, R., Ingraham, L., Darnell, D., Whiteside, L., & Van Eaton, E. (2015). Technology-Enhanced Stepped Collaborative Care Targeting Posttraumatic Stress Disorder and Comorbidity After Injury: A Randomized Controlled Trial. *Journal of Traumatic Stress*, *28*(5), 391–400. Scopus. 10.1002/jts.22041.

Zhao, K., & So, H.-C. (2019). Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE Journal of Biomedical and Health Informatics*, *23*(3), 1304–1315. Scopus. 10.1109/JBHI.2018.2856535.

Zhuang, X., Rozgić, V., Crystal, M., & Marx, B. P. (2014). Improving speech-based PTSD detection via multi-view learning. *IEEE Spoken Language Technology Workshop (SLT), 2014*, 260–265. https://doi.org/10.1109/SLT.2014.7078584

Appendix 3

[III]

M. Bertl, P. Ross, and D. Draheim. Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors. *Health Policy and Technology*, 2023

Health Policy and Technology 12 (2023) 100748

Contents lists available at ScienceDirect

Health Policy and Technology

journal homepage: www.elsevier.com/locate/hlpt

Original Article/Research

Systematic AI Support for Decision-Making in the Healthcare Sector: Obstacles and Success Factors

Markus Bertl [a,*], Peeter Ross [a], Dirk Draheim [b]

[a] Department of Health Technologies, Tallinn University of Technology, Akadeemia tee 15a, Tallinn, 12616, Estonia
[b] Information Systems Group, Department of Software Science, Tallinn University of Technology, Akadeemia tee 15a, Tallinn, 12616, Estonia

ARTICLE INFO

Keywords:
Decision support systems
Healthcare information systems
Health informatics
Delivery of health care
Artificial intelligence (AI)
Machine learning (ML)
Decision-making
GAIA-X, e-health, Digital health

ABSTRACT

Background: Currently, health care is expert-centric, especially with regard to decision-making. Innovations such as artificial intelligence (AI) or interconnected electronic health records (EHRs) suffer from low adoption rates. In the rare cases of technically successful implementation, they often result in inefficient or error-prone processes.
Aim & Methods: This paper explores the state of the art in AI-based digital decision support systems (DDSSs). To overcome the low adoption rates, we propose a systematic strategy for bringing DDSS research into clinical practice based on a design science approach. DDSSs can transform health care to be more innovative, patient-centric, accurate and efficient. We contribute by providing a framework for the successful development, evaluation and analysis of systems for AI-based decision-making. This framework is then evaluated using focus group interviews.
Results: Centred around our framework, we define a systematic approach for the use of AI in health care. Our systematic AI support approach highlights essential perspectives on DDSSs for systematic development and analysis. The aim is to develop and promote robust and optimal practices for clinical investigation and evaluation of DDSS in order to encourage their adoption rates. The framework contains the following dimensions: disease, data, technology, user groups, validation, decision and maturity.
Conclusion: DDSSs focusing on only one framework dimension are generally not successful; therefore, we propose to consider each framework dimension during analysis, design, implementation and evaluation so as to raise the number of DDSSs used in clinical practice.
Public Interest Summary: The digital transformation of the healthcare sector creates the potential for the sector to be more accurate, efficient and patient-centric using AI, or so-called digital decision support systems. In this research, we explore why these systems are needed and how they can be successfully implemented in clinical practice. For this, we propose a systematic approach based on our conceptual framework. Against this background, we present our vision for further advancing these technologies. We see our systematic AI support as a primary driver, with the possibility to facilitate the much-needed breakthrough of decision support systems in health care.

1. Introduction

Innovations in the health sector are of pivotal importance for today's societies. Consequently, e-health, recently also referred to as digital health, is an essential topic in eGovernment [1] and smart city research [2,3]. Smart, sustainable cities are data-driven. However, digital transformation in the healthcare sector requires changing clinical workflows established a long time ago [4]. This change towards an improvement in quality is influenced by policies, health management and (clinical) care [5]. In all of the three mentioned areas, decisions need to be made. These decisions have a direct impact on patients' lives, on caregivers as well as on society at large. As one example, 250,000 Americans die from medical errors each year [6]. A medical error costs hospitals $939 on average, totalling $1 billion for US hospitals alone [7]. According to [8], drivers of poor medical care can be grouped into (i) "money, finance, and organization"; (ii) "knowledge, beliefs, assumptions, bias, and uncertainty"; and (iii) "power and human relationships". Additionally, the absence of evidence for clinicians as well as biased research were identified as reasons for errors in medical decisions [8]. According to the survey in [9], 37% of healthcare organisations lack the data they need

* Corresponding author.
E-mail address: mbertl@taltech.ee (M. Bertl).

https://doi.org/10.1016/j.hlpt.2023.100748

Available online 27 April 2023

for decision-making. On the other hand, the creation of health-related data is rising year by year. We can assume that some of the evidence needed may already be present in current hospital information systems. However, due to legacy data silos caused by a lack of interoperability and data connectivity, these data are not always accessible for medical decision-making. We also hypothesise that the issue of interoperability is not merely a lack of common terminologies and archetypes, messaging or database formats and technical connectivity. Rather, the problem originates in the inability of the information systems to adapt to a continuous need for learning – for the flexible readjusting of ontologies of cognition and remaining interoperable at the same time. Therefore, the question of the trade-off between rigid standards and faster learning arises.

If timely, high-quality data are provided, we see high potential for AI-driven digital decision support systems (DDSS). The Cambridge Dictionary defines AI as "the study of how to produce computers that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, solve problems, and learn" [10].

Sauter defines DDSSs as "computer-based systems that bring together information from various sources, assist in the organization and analysis of information and facilitate the evaluation of assumptions underlying the use of specific models" [11]. In the medical domain, DDSSs can optimise patient care by improving decisions in the aforementioned areas that drive poor medical care. DDSSs, therefore, play a vital role in making our cities, regions and communities smart and sustainable. Ironically, instead of improving the situation, AI and DDSSs currently play a niche role, even introducing new kinds of medical decision-making problems. In this research, we explore this role and provide a vision ensuring that DDSSs fulfil the expectations for making health care safer, better and more efficient. We propose that the key to these goals is a systematic approach, which we call systematic AI support.

2. Background: The low adoption rates of AI-based DDSS in health care

Recent studies suggest problems with AI-based DDSSs [12–14] despite their often cited potential. According to Heeks [15], up to 85% of health IT projects encounter some kind of failure. He assumes that traditional, structured development methodologies are one reason for such failure. Another problem is data alone. We lack structured electronic documentation, centralised and connected EHRs and have high data protection standards, making the collection and reuse of data difficult. Not only do these problems slow down the adoption of AI-based DDSSs, but low data quality also hinders administration of care based on evidence-based practice.

Another factor negatively influencing DDSSs adoption is low user acceptance [12] caused by low IT literacy, lack of training and support of staff to use e-health systems, lack of time as well as missing funding for health IT [13]. Furthermore, systems are difficult to use because their integration with clinical workflows is often unsuccessful [16]. In the end, health professionals end up with dysfunctional systems that frustrate users and ironically lead to errors and avoidance of DDSSs altogether [17,18]. Additional negative factors include obscured responsibilities between computers and humans leading to liability discussions and threats to clinicians' independence [19].

In radiology, standardised digital imaging and structured reporting have existed for many years, meaning that a large foundation of labelled data is available to train AI algorithms. The anomaly detection algorithms developed are even better than humans in some cases. Still, AI has not replaced radiologists quite yet, nor has it even found its way into radiologists' routine daily practice. The reasons for this are the underestimation of the number of variables that influence meaningful advice, the need for iterative communication with colleagues and the consideration of prior health and medical data for decision-making. Anomaly alone does not lead to meaningful reports or treatment [20]. Furthermore, not every anomaly is a reason for disease. AI alone brings up too many insignificant abnormalities. This forces radiologists to investigate AI-selected anomalies, leading to unnecessarily high time consumption. Additionally, current AI systems do not suit clinical workflows.

There are rare examples of DDSSs in production, some even with high adoption rates. One is EBMeDS from Duodecim [21]. EBMeDS works on top of Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and uses a rule-based approach to assist with clinical guidelines and provide clinical reminders and drug assistants [22]. EBMeDS includes an organisational solution for creating a feedback loop for algorithm validation, which we assume as one success factor for smoother implementation. Others are functionality for the organisation to measure DDSS performance and thereby control the quality of input data and user habits. Current research projects mostly lack these functionalities [14]. EBMeDS supports 3646 evidence links, 21,762 drug interactions and 988 custom scripts in 13 languages [23]. Another popular example is IBM's Watson Health Platform, consisting of many modules that help analyse diagnostics, drug interactions, radiological images, oncological treatment or administrative tasks. Although used in production, some researchers are questioning whether it brings real value [24]. Currently, most tools in use only assist medical decision-making by providing information from scholarly publications. Examples are Wolters Kluwers UpToDate [25] or Elseviers ClinicalKey [26].

We see successful DDSSs as socio-technical learning systems, which have been trained by providing real-life feedback over a longer period. This enables the systems to overcome the difficulties encountered during their immaturity phase. Systems deployed in multi-institutional setups complicate fast feedback with high accuracy. Therefore, systems like EBMeDS are usually deployed internally for one institution, where DDSSs start to influence the feedback on organisational behaviour and data quality.

3. Comparison with previous work

Many frameworks for healthcare technology have been published, which have been reviewed and analysed by Greenhalgh et al. [27]. As stated by Greenes et al. [28], it is challenging to find a single overarching model that covers all the aspects of healthcare technology in sufficient detail to remain useful. Therefore, we see the need for multiple models or frameworks for various aspects of this complex domain. None of the currently published models deals specifically with AI-based decision support for health care. Popular frameworks in health informatics, like NASSS [27], do not address the practical shortcomings of DDSS development so far. Using our framework with the novel healthcare technology maturity levels, we highlight the necessity of advancing maturity levels to the highest level in order to enable use of a system in real clinical work. This is especially important because the development process is long and systems developed in isolation under 'textbook' conditions (e.g. algorithms with lower maturity levels) have high potential for non-adoption.

4. Methods

Our systematic AI support approach is visualised on the basis of a theoretical framework. For the development of our framework, we followed the design science research paradigm proposed by Hevner et al. [29]. Design science is used as a systematic methodology for developing and evaluating a novel design artefact (technology, framework, etc.) that copes with a real-world problem. According to Hevner et al. [29], design science research is the process of creating a purposeful artefact for a specific domain with a comprehensive evaluation. The designed artefact needs to solve a specific problem that is rigorously defined, formally represented, and coherent and internally consistent. Design science has been widely accepted as an information systems research

method [30,31] and is a particularly good fit for our research purpose, since it helps to address both the role of IT artefacts in information system research [32] and the low level of professional relevance of many studies [33] on DDSSs.

We apply the design science approach to DDSSs by developing a novel framework for raising the adoption rates of AI-based DDSSs. The resulting artefact of our design process – our framework – is further described in Section 5. The evaluation based on a focus group interview is laid out in Section 6.

5. Results: A framework for improving AI-based DDSS adoption rates

Fig. 1 presents our framework, designed based on terminology extracted from two previous systematic literature reviews [14,34] and our own observations. The terminology was extracted from 80 journal articles on DDSSs or AI in psychiatry using thematic analysis [35] and narrative synthesis [36].

The framework specifies key variables influencing DDSSs and provides scope for defining research problems, conducting reviews, evaluating a solution or benchmarking solutions. It allows the systematic analysis of the steps needed for successful DDSSs adoption. Our framework's terminology also serves as a common language to bridge barriers in interdisciplinary medical informatics. Since health care is a complex, adaptive system [37], we cannot develop new approaches or study issues based on isolated entities. New systems change the environment in which they are deployed and therefore complex feedback loops emerge between all involved systems. Each successful approach needs to be able to cope with such complexity. Based on the principle of separation of concerns, our framework can be used to divide the overall problem of DDSSs into smaller parts for investigation. It draws attention to distinct aspects of DDSS development, thereby supporting the coordination of profoundly specialised expertise. Our framework is not a static tool; as the environment changes, each dimension needs to be re-evaluated.

DDSSs can fulfil many tasks. Therefore, subcategories are needed for a more evident scope. DDSSs can be classified into two categories based on their primary purpose:

- Data entering decision support. This covers everything supporting the data entering information systems, e.g. taxonomies (ICD-10, SNOMED-CT, etc.), template tools, structuring or reporting tools.
- Decision support based on collected data offers support based on data reuse. We divide these systems further by their use into six areas (adapted from Ross et al. [38]):
  - Improved data usage and visualisation aim to aggregate and visualise data based on open data repositories, EHRs and personalised health records.
  - Health management allows personalised care management to improve disease prevention, screening and case management.
  - Patient monitoring involves summarised methods of analysing the health status of individuals to visualise treatment outcomes or the need for intervention. This also includes telemedicine solutions, patient diaries (incl. lifestyle data), prescription alerts or treatment adherence monitoring.
  - Clinical decision support systems provide support for diagnoses or treatment decisions.
  - Scientific gene research. Research based on genomic data.
  - Public health analysis tools aim to aggregate and visualise data to find health trends for a population and uncover and support health promotion programmes.

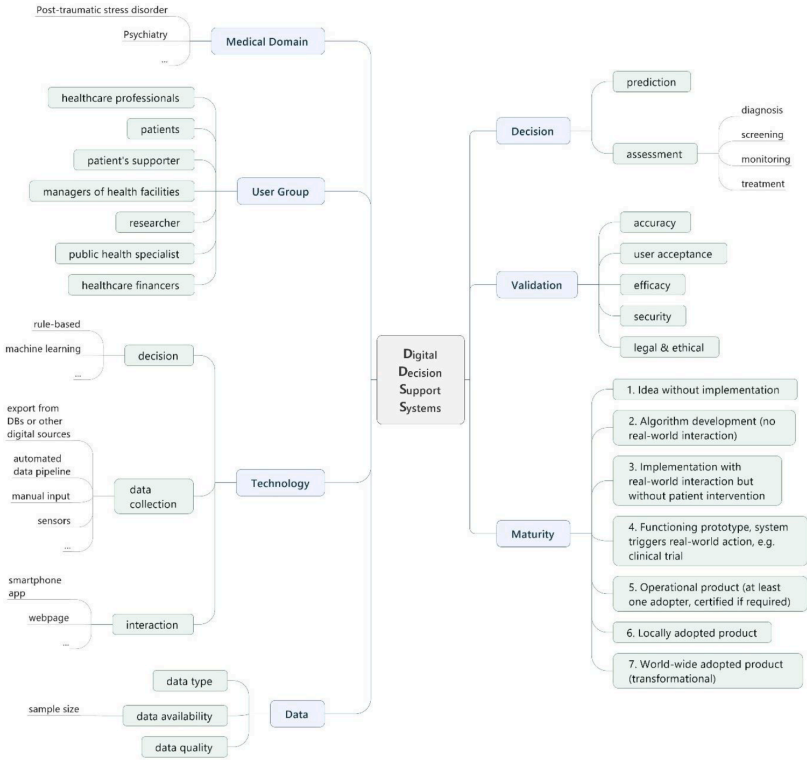


Fig. 1. Framework for systematic AI support.

DDSSs can also be differentiated from a workflow point of view as follows:

- *DDSS with a human in the loop* supports tasks. However, the end decision is always made or reviewed by humans.
- *DDSS without a human in the loop*, or automated decision-making, acts completely autonomously without the need for any user approval or review.

Every DDSS category requires unique framework dimensions. Although the framework applies to all kinds of DDSS, we focus our vision for the future on clinical decision support, patient monitoring, health management and improved data use and visualisation. Therefore, we facilitate every system that helps patients, caregivers and clinicians in making health-related decisions more effective. The following subsections describe the framework dimensions in more detail.

5.1. The data dimension

The input data dimension defines the information needed by a DDSS to function. Data power the decision technology and therefore define the core outcomes of the system. They are a key component for the success of DDSSs. Not surprisingly, data quality needs to be checked and ensured to avoid bias and undesirable results. Recent studies even suggest that data quality is a crucial success factor for AI in health care [39,40]. Therefore, a lot of thought needs to be put into DDSS input data.

Possible input data could be structured, such as metadata (e.g. sociodemographic information), numerical data, coded data (e.g. ICD-10) as well as semi-structured information (e.g. patient records), unstructured information such as free text, or medical graphs (e.g. ECG, EEG) and images (e.g. X-ray, ultrasound). It needs to be ensured that the data used for developing a decision support system are also available if the system should be used in production later on. This is crucial because many research articles only propose systems based on available datasets. These data might not be available in a real-world setting. Systems developed in isolation under those 'textbook' conditions have a high potential for failure.

Due to the lack of centralised EHR systems, interoperability and standardisation, data reuse for DDSSs is often difficult to achieve [14]. This is especially true nationwide. Even if the legal and organisational challenges to accessing these data can be solved, decentralised storage of healthcare data and the need for integration of several data sources throw up additional technical barriers for DDSSs. One example of how data integration could be handled is through the use of grids and peer-to-peer networks. They can be used to integrate distributed, often heterogenous data sources at geographically distributed sites [41]. Additionally, cloud computing offers many possibilities, provided that ways to ensure national data security standards (like HIPAA, PIPEDA, or GDPR) are found [42]. In Europe, recent developments around GAIA-X, the European Data Platform, also attempt to tackle these problems [43,44]. GAIA-X was initially launched as a trusted provider of next-generation data infrastructure, meeting the highest standards in terms of data protection and data sovereignty (providing data owners complete control over their data). The GAIA-X ecosystem enables data linkage through federation services for use at a regional, national or European level. This is possible because GAIA-X provides and enforces a common set of policy rules and an architecture of standards of interconnection. This could be a game-changer for the healthcare sector. In further research, we will demonstrate how GAIA-X can be used for DDSSs in health care.

Nevertheless, most data in EHRs are created by humans and therefore reflect their social and cultural context, cognitive biases, experience and emotions, sometimes even their beliefs [8]. Quality assurance and curation of such datasets is difficult and time-consuming. Additionally, in areas like mental health, decision-making is often more disease-focused than symptom-focused. The curative medical model suggests treatment based on eradication or slowing down disease progression [45]. This means that treatment success is measured in disease-related terms, such as tumour size or rate of survival. Outward manifestations such as quality of life or the patient's subjective feeling are often of a lower priority. This reflects on the data used for research. Diagnoses are already an aggregation of data. Symptoms leading to a diagnosis are not always included in datasets and therefore not included by DDSS decision-making. This adds another level of bias and makes feedback loops for the correction of wrong diagnoses difficult.

One way of avoiding these biases is unobtrusive data collection. Examples are sensors in wearable devices, such as smartphones; some sensors can even be printed on the skin or in textiles [46]. While diagnosis data in EHRs are aggregations of physiological measurements and symptoms, capturing and evaluating these data directly in an unobtrusive way allows for a less biased, symptom-focused (and therefore also patient-focused) approach to decision-making. Unobtrusive data collection is more robust and less labour-intensive, limits research bias (Hawthorne effect) and has lower associated costs. Rather than having data collected only at doctor's visits, this allows a holistic view of current problems by capturing signals more frequently.

5.2. The technology dimension

The technology dimension describes how DDSSs are implemented based on three sub-dimensions:

- *Decision technology* is the algorithm that powers the decision-making. It can be understood as the *brain* of a DDSS. A decision can be empowered by AI, statistics, rules or mixed methods. AI as decision technology is already mature and performs well, as indicated by the success of global data-driven companies like Amazon, Google and Facebook. This is why we argue that the poor performance of DDSSs in health care is not so much connected to decision technology but rather to data (5.1), data collection, user interaction (5.3) and validation (5.4). However, in health care especially, AI algorithms should consider reproducibility and explainability.
- *Interaction technology* describes the ways of interacting with systems, user groups or the clinical process. Examples include APIs, graphical user interfaces (e.g. websites, mobile apps) or sensory input such as conversational interfaces (chatbots). If the interaction does not fit its user's needs (e.g. disruptive alerts), it becomes a major barrier. Additionally, computer literacy needs to be taken into consideration when planning the interaction.
- *Data collection technology* defines how the data described in the input data dimension are gathered. Examples are sensors, questionnaires or chatbots. As described above, we hypothesise that the future of data collection will be a combination of high-quality structured sources such as centralised or federated EHRs as well as unobtrusive data that can be used as interaction and collection technology for decision-making (e.g. virtual agents collecting speech data).

5.3. The user group dimension

This dimension captures the user groups interacting with DDSSs in the clinical process. User groups play a vital role in the adoption of DDSSs. In general, physicians tend to distrust AI systems regardless of their output's accuracy [20]. The different user groups need to be analysed and should never be neglected in the conception, development and testing phase, so as to raise user acceptance. This also ensures that IT solutions fit the clinical workflow and that the resulting application is usable in production. *Re*-engineered workflows by IT professionals instead of clinicians can lead to the opposite effect. DDSSs can negatively disrupt the clinical workflow, especially if they are standalone systems or are not integrated into the IT landscape. Such disrupted workflows can have adverse effects, such as increased cognitive effort and higher time consumption [47]. Yet, this does not mean that clinical

workflows should not be changed. On the contrary, the introduction of new technology to old processes is another reason why AI cannot fulfil its potential. The core concept of medicine has not changed in the last 20 years or more; however, the world around it has. Therefore, processes need to be adapted to the current situation. Process adaptations often result in an identity crisis of healthcare workers and are experienced more often as barriers and threats to the professional identity than as support. Questions concerning boundaries arise. What is the role of doctors or the responsibility of IT systems? Other domains clearly show that new technologies need new workflows. As an example, successful online retailer Amazon operates entirely unlike any other retailer. We claim that such positive disruption is also needed in health care in order to leverage AI's full potential. Here, apart from the clinical process, possible harmful long-term effects need to be continuously addressed. As soon as users adapt to the safety net provided by DDSSs, they might start to make more mistakes without them [48,49]. On the other hand, users learn by using DDSSs; previously useful notifications can start to feel disruptive [50].

The traditional use of DDSSs, which has been to provide direct alerts and guidance to physicians in making treatment decisions for a particular patient, has been extended. These further uses include recommendations for improving workflow, cost savings, prevention as well as public health, research and health policy decisions. Similarly, DDSS user groups have expanded. In addition to physicians and nurses, decision support can assist hospital managers in business-critical decisions, insurance agencies in planning funding and researchers in identifying patients eligible for research based on specific criteria. The number of decision support systems for patients is increasing rapidly with the wider deployment of Personal Health Records (PHR).

The features of DDSSs and thus the development should be based on the needs and expectations of a specific user group. Based on the recommendation or analytics offered by the DDSS, as well as the time-critical nature of decision-making, we propose to group DDSS users as follows:

- Healthcare professionals
- Patients
- Patient's supporters
- Managers of healthcare facilities
- Researchers
- Public health specialists
- Healthcare financers

The decision-making of healthcare professionals at an appointment or at a bedside is time-critical and requires high-quality operational data from the EMR and/or EHR. In general, patient support is person-specific and relatively time-dependant, but prevention and genome-based algorithms can provide advice in the longer term. The second group consists of researchers, public health professionals and health financiers, for whom decision support uses secondary, often aggregated and non-time-critical data. Healthcare managers need both short-term data, such as hourly or daily hospital bed availability alerts, and aggregated data for advice on staffing or budget planning.

Targeting user groups more specifically, we aim to develop and promote robust and optimal practices for clinical investigation and evaluation of DDSSs to encourage their adoption rates. Similarly, an increase in the user-orientated development of algorithms and applications in turn increases trust in DDSSs.

5.4. The validation dimension

Validation describes the measurement of success of DDSSs, categorised into four sub-dimensions:

- *Accuracy* describes evaluation based on how many right or wrong decisions a system makes. Example measurements are algorithmic accuracy, area under the curve (AUC) values [51], F1 scores, recall (sensitivity), precision, and specificity [52].
- *User acceptance* is about evaluating the perceived usefulness and perceived ease of use of DDSSs [53].
- *Efficacy* evaluates the impact of systems based on potential benefits.

Security needs to be an integral part of DDSS development to ensure the adoptability of the technology later on. The importance of enhancing security in the healthcare industry has been highlighted by recent studies on the increase of cyber-attacks on digital health infrastructure [54,55].

- The *legal & ethics* sub-dimension describes the evaluation of legal regulations and ethical considerations (e.g. medical device regulations, data protection regulations, responsibility and reliability of the DDSS output and its relation to the human decision). This dimension was identified as a key success factor for AI adoption [56].

For successful DDSS adoption, evaluation needs to be standardised to ensure trust. We propose the use of unified benchmark datasets in order to be able to compare the cognition performance of different systems. However, validation of AI based on accuracy, which evaluates the decision technology dimension, is not enough. Additionally, evaluations for user interaction technology (user acceptance and efficacy) and data collection technology (accuracy, legal/compliance) need to be carried out to ensure that the system meets user expectations. For DDSSs with a maturity level of 3 or higher (see 5.7), evaluation categories need to be investigated not only for each dimension of our framework separately but also for the entire system. Each dimension itself might fulfil all the legal requirements; put together, however, the system still might not be legally compliant.

5.5. The medical domain dimension

The *medical domain dimension* describes the particular illness, medical condition or health area for which DDSSs can be applied. Even though the examples in the presented framework contain only two psychiatric diseases, in the design and implementation of DDSSs, the clinical domain of planned systems should be considered and relevant clinical specialists involved in the process. One crucial success factor is to find the right granularity in this dimension. If the dimension is too specific, DDSSs tend not to consider all the necessary variables. If it is too general, DDSSs do not benefit the users. As an example, we show this based on DDSSs in radiology. A too broadly defined DDSS might only say whether a radiological image is abnormal. However, this would not be of much value to a physician since they still need to look at the abnormality, interpret it within the patient's medical context and compile a meaningful report based thereon. A too narrowly defined DDSS might only look for one specific finding, such as lung nodules, but might miss other pathologies present in the radiological image. EBMeDS is one example where algorithms are divided between disease category and medical domain for better results.

In other terms, this dimension deals with the necessary medical background knowledge about the domain. This is important not only to provide the right domain context for the system (e.g. what decision a system should support in order to improve the clinical process), but also because it deals with the way in which knowledge is transferred from the real world into the DDSS. This might be based on manually created rules, like in EBMeDS, or machine-learning or deep-learning from certain data. Even though rules can be generated automatically from data and AI can learn unsupervised, the current narrow AI is still goal orientated and only capable of providing specific tasks. As long as we don't speak about general AI, which is still far away, a domain expert is needed to select the right knowledge source for the system to learn from. Otherwise, DDSS research might yield good results at low maturity levels (e.g. high accuracy values), but systems will potentially still not perform well in a

production setting because assumptions about the knowledge base (selected rules or used training data) do not accurately reflect the clinical reality. In other words, the output of current DDSSs is mostly based on shallow knowledge. However, for most cases, the deep knowledge of medical professionals is still required for verification, for consideration of additional variables not contained in the knowledge source of the DDSS (like social-economic status or other prior health and medical data) and for communication.

5.6. The decision dimension

We classify decisions into *prediction,* where the system outputs a risk score based on the likelihood of getting a disease, and *assessment*, where the patient is already sick (knowingly or unknowingly). *Assessment* contains the subcategories *diagnosis* (for individuals with symptoms or suspicion of illness), *screening* (for individuals without specific symptoms), *monitoring* (evaluates symptom severity or treatment progress) and *treatment* (recommendation, automated or manual intervention concerning care or therapy).

5.7. The maturity dimension

To describe the maturity of DDSSs, we suggest seven levels that provide a common understanding of the development status, transition and measurement of DDSS research progress (see Table 1). Additionally, they serve as a risk management tool for implementation considerations. Our maturity level approach is derived from the technology readiness levels from NASA [57]. This scale has been adopted by many institutions like the European Union to better describe the expectations and status of research results [58]. While preserving the initial idea of assessing the maturity of technology for production use, we changed the terminology and scales to fit the digital health research.

When using our framework in combination with the maturity levels, it is essential to note that not all dimensions are necessarily present in detail at each maturity level. As the level gets higher, more dimensions should be described to enable successful system application in real-world practice.

Currently, AI in health care has low maturity [14]. The reasons for this are evident in the data, user group and validation dimensions. These are the dimensions that are often not clearly defined in lower-level DDSSs. Research focusing on pure algorithmic development of AI systems usually takes certain datasets without validating whether the data are available in clinical practice, does not consider processes in health care or user groups and validates only against standard AI metrics such as accuracy [14,34]. This results in theoretically sound algorithms that are, unfortunately, not ready for higher maturity DDSSs. We suggest a more comprehensive approach using our framework throughout the development process to ensure that DDSSs can be brought to higher maturity. As a limitation, it is important to note that our framework does not consider commercial perspectives of decision support.

Table 1
Maturity levels.

Level	Description
1	Idea without implementation
2	Implementation without real-world interaction (algorithm development)
3	Implementation with real-world interaction but without patient intervention (no real intervention on a patient takes part based on the output of the DDSS)
4	Fully functioning prototype, system triggers real-world action (e.g. clinical trial)
5	Operational product (at least one adopter, certified if required)
6	Locally adopted product
7	Worldwide adopted product (transformational)

6. Evaluation

We used a focus group interview to evaluate and receive in-depth feedback on the usefulness of our framework. For the focus group, we selected international experts in DDSSs or one or more of our framework dimensions. Nine experts from four countries (Austria, Estonia, Sweden, Ukraine) of the following professions took part in the interview:

- Programme Director – Business Information Technology
- Lawyer – Subject Matter Expert for data privacy in health care
- Social Scientist
- Software Architect, Subject Matter Expert for secondary use of health data
- Expert for Interoperability in health care
- Data Integrity & Transparency Expert
- CEO of a digital health company
- Lead Data Scientist
- Professor, Medical Doctor

They received the framework with a description of the dimensions two weeks prior to the session. The focus group session was conducted remotely via the MS Teams platform to accommodate the locations of the experts. The interview followed a semi-structured format, with moderation by the lead author. In the first 15 min, the authors explained the protocol for the focus group interview and the participants were introduced to one another. Next, the framework was presented to make sure that all participants had a common understanding of the discussion topics. The rest of the session consisted of an interactive discussion around the usefulness and application of our designed framework. We recorded the 90-minute session. Afterwards, the authors coded the themes of the recording individually and aggregated the emerging topics. Clustering was based on inductive thematic analysis according to Brown and Clark [35]. The information was compared, reflected on and condensed by the authors together until a consensus was reached.

The main observation was that all participants saw value in the framework for analysing DDSS. In regard to the usefulness of the framework in the strategic planning and development of DDSSs, the majority of participants would require more detailed sub-dimensions of the framework. One example is the legal dimension, which could be expanded to cover European Union regulations like the General Data Protection Regulation (GDPR) or the Medical Device Regulation (MDR). For the use of the framework in the industry, the experts advised adding a dimension around commercial and financial aspects. The general recommendation was to provide a clear scope and use groups for the framework from the above points. The results of our focus group and the implications on the framework are discussed in Section 7.

7. Discussion, limitations and further research

Although our framework was developed as a conceptual framework to analyse and aggregate current literature on DDSSs, possible applications are far more widespread. From the initial idea of a DDSS to the entire development cycle, our framework can be leveraged as boilerplate by product owners or decision-makers. Used as a blueprint, it gives guidance on which areas (framework dimensions) need to be considered. This is especially important because DDSS development for health care is an interdisciplinary research area. It requires IT knowledge as well as domain knowledge in health care. The suggested framework helps make this transparent to the two groups and to support a multidisciplinary approach. Our framework is developed based on literature and expert opinions from technology and healthcare perspectives. It is one of the first to consider both domains and therefore has potential to accelerate not only DDSS adoption but also AI and ML adoption.

Additional use cases concern the evaluation and maturity assessment of DDSSs. Using a standardised method of assessment of current DDSSs helps make the various products in this vast growing market

comparable.

Our framework does not include a financial or commercial domain. We see this domain as separate from actual DDSS development and therefore did not include further details on this. Additionally, the framework only deals with the general components of DDSSs and DDSS evaluation. More detailed subdimensions could further enrich our dimensions. One example is the legal dimension, where a specific subdimension could serve as a checklist to highlight what rules and regulations need to be taken into consideration when developing a DDSS. More detailed sub-dimensions tend to become increasingly country-specific. For now, we decided not to go into further detail for this research in order to make it applicable worldwide, though we acknowledge that this is an area for further research.

Another area of further research is the application of the framework to develop and evaluate decision support systems. We are planning to use the framework to expand our work on automated rule generation from health insurance claims data [59] into a decision support system. This will produce further evidence as to the usefulness of a systematic approach in raising DDSSs adoption rates.

8. Conclusion

The healthcare sector is currently facing many challenges, such as ever-increasing complexity and information overload. This results in inefficient decision-making, which may cause errors. State-of-the-art DDSSs can transform the healthcare sector into a more efficient and patient-centric operation. Currently, such systems have only low adoption rates, and success stories about AI bringing real value to clinicians or patients are rare. Our systematic AI support approach aims to bring DDSSs into production. By elaborating on essential perspectives from both health care and IT, we created a framework for the systematic development and analysis of DDSSs. Currently, most DDSS research focuses on either the medical or the technical domain. Since digital health is an interdisciplinary subject, investigating DDSSs solely from a technical or medical perspective is insufficient. Therefore, DDSSs that focus on only one framework dimension are generally unsuccessful. The development of AI algorithms without any medical knowledge and context or those based on inadequate datasets are especially likely to bring no real-world clinical value and are set up to fail. We propose to consider each framework dimension during architecture, development and evaluation. Similarly, the framework can function as a means to divide DDSS development based on expertise, allowing us to bring in experts from several subject matters while still maintaining a comprehensive approach. This ensures that all necessary features of a DDSS are investigated by experts for each dimension, from both a technical and an organisational perspective. Therefore, we see the proposed systematic AI support as a major driver for the much-needed breakthrough of DDSSs in health care.

Funding

The researchers did not receive funding for this research.

Ethical approval

Not required.

Patient consent

Not required.

CRediT authorship contribution statement

Markus Bertl: Conceptualization, Methodology, Investigation, Resources, Writing – original draft. **Peeter Ross:** Conceptualization, Resources, Methodology, Writing – review & editing, Supervision. **Dirk Draheim:** Conceptualization, Resources, Methodology, Writing – review & editing, Supervision.

Declaration of Competing Interest

No conflict of interest to declare.

References

[1] Kovac ME. Health demystified: an e-government showcase. Computer. 2014 Oct;47(10):34–42.
[2] Alexopoulos C., Pereira G.V., Charalabidis Y., Madrid L. A taxonomy of smart cities initiatives. In: Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2021 May 7]. p. 281–90. (ICEGOV2019). Available from: 10.1145/3326365.3326402.
[3] Pramanik MI, Lau RYK, Demirkan H, MdAK Azad. Smart health: big data enabled health paradigm within smart cities. Expert Syst Appl 2017;87:370–83.
[4] Warraich HJ, Califf RM, Krumholz HM. The digital transformation of medicine can revitalize the patient-clinician relationship. Npj Digit Med 2018;1(1):1–3.
[5] Reid PP, Compton WD, Grossman JH, Fanjiang G. Building a better delivery system: a new engineering/health care partnership. National Academies Press (US); 2005 [Internet][cited 2021 Feb 23]. Available from, https://www.ncbi.nlm.nih.gov/books/NBK22878/.
[6] Makary MA, Daniel M. Medical error—The third leading cause of death in the US. BMJ 2016;353 [Internet][cited 2021 May 12]Available from, https://www.bmj.com/content/353/bmj.i2139.
[7] David G, Gunnarsson CL, Waters HC, Horblyuk R, Kaplan HS. Economic measurement of medical errors using a hospital claims database. Value Health 2013;16(2):305–10.
[8] Saini V, Garcia-Armesto S, Klemperer D, Paris V, Elshaug AG, Brownlee S, et al. Drivers of poor medical care. The Lancet 2017;390(10090):178–90.
[9] Cortada J, Gordon D, Lenihan B. The value of analytics in healthcare - from insights to outcomes [Internet]. IBM Institute for Bus Value 2012 [cited 2021 Jan 14] Available from, https://www.ibm.com/downloads/cas/NJA9K0DV.
[10] Artificial Intelligence. In: Cambridge international dictionary of English. Cambridge: Cambridge University Press; 2014 [Internet][cited 2021 Feb 15] Available from, https://dictionary.cambridge.org/de/worterbuch/englisch/artificial-intelligence.
[11] Sauter VL. Decision support systems for business intelligence. John Wiley & Sons; 1997, 618.
[12] Hennemann S, Beutel ME, Zwerenz R. Ready for eHealth? Health professionals' acceptance and adoption of eHealth interventions in inpatient routine care. J Health Commun 2017;22(3):274–84.
[13] Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010;10(1):231.
[14] Bertl M, Ross P, Draheim D. A survey on AI and decision support systems in psychiatry – Uncovering a dilemma. Expert Syst Appl 2022;202:117464.
[15] Heeks R. Health information systems: failure, success and improvisation. Int J Med Inf 2006;75(2):125–37.
[16] Kellermann AL, Jones SS. What it will take to achieve the as-yet-unfulfilled promises of health information technology. Health Aff (Millwood) 2013;32(1):63–8.
[17] Yen PY, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc 2012;19(3):413–22.
[18] Pizziferri L, Kittler AF, Volk LA, Honour MM, Gupta S, Wang S, et al. Primary care physician time utilization before and after implementation of an electronic health record: a time-motion study. J Biomed Inform 2005;38(3):176–88.
[19] Varonen H, Kortteisto T, Kaila M. What may help or hinder the implementation of computerized decision support systems (CDSSs): a focus group study with physicians. Fam Pract 2008;25:162–7.
[20] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. Npj Digit Med 2021;4(1):1–8.
[21] Duodecim | EBMEDS (Evidence-Based Medicine electronic Decision Support) [Internet]. [cited 2021 Mar 2]. Available from: https://www.ebmeds.org/en/.
[22] Duodecim Medical Publications Ltd. EBMEDS White Paper [Internet]. 2020 [cited 2021 Jan 17]. Available from: https://www.ebmeds.org/wp-content/uploads/sites/16/2020/10/WhitePaper_2020-1.pdf.
[23] Statistics – Duodecim | EBMEDS [Internet]. DUODECIM - EBMEDS clinical decision support. 2020 [cited 2021 Jan 18]. Available from: https://www.ebmeds.org/en/materials/statistics/.
[24] Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. IEEE Spectr 2019;56(4):24–31.
[25] Evidence-Based Clinical Decision Support at the Point of Care | UpToDate [Internet]. [cited 2021 Mar 2]. Available from: https://www.uptodate.com/home.
[26] ClinicalKey - Lead with Answers [Internet]. [cited 2021 Mar 2]. Available from: https://www.clinicalkey.com/#!/.
[27] Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e8775.

[28] 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.
[29] Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q 2004;28(1):75–105.
[30] Indulska M, In Recker JDesign science in IS research : a literature analysis. Information systems foundations: the role of design science. ANU Press; 2008. p. 285–302.
[31] Arnott D, Pervan G. A critical analysis of decision support systems research revisited: the rise of design science. J Inf Technol 2014;29(4):269–93.
[32] Orlikowski WJ, Iacono CS. Research commentary: desperately seeking the "IT" in IT research—a call to theorizing the IT artifact. Inf Syst Res 2001;12(2):121–34.
[33] Benbasat I, Zmud RW. Empirical research in information systems: the practice of relevance. MIS Q 1999;23(1):3–16.
[34] Bertl M, Metsallik J, Ross P. A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. Front Psychiatry 2022;13.
[35] Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77–101.
[36] Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme 2006.
[37] McBride K, Draheim D. On complex adaptive systems and electronic government: a proposed theoretical approach for electronic government studies. Electron J E-Gov 2020;18(1). pp43-53-pp43-53.
[38] Ross P. Feasibility study for the development of digital decision support systems for personalised medicine. Estonian Ministry of Social Affairs; 2015.
[39] Faraj S, Pachidi S, Sayegh K. Working and organizing in the age of the learning algorithm. Inf Organ 2018;28(1):62–70.
[40] Wirtz BW, Weyerer JC, Geyer C. Artificial intelligence and the public sector—applications and challenges. Int J Public Adm 2019;42(7):596–615.
[41] Comito C, Talia D. GDIS: a service-based architecture for data integration on grids. In: Meersman R, Tari Z, Corsaro A, editors. On the move to meaningful internet systems 2004: otm 2004 workshops. Berlin, Heidelberg: Springer; 2004. p. 88–98 (Lecture Notes in Computer Science).
[42] Fernández-Cardeñosa G, de la Torre-Díez I, López-Coronado M, Rodrigues JJPC. Analysis of cloud-based solutions on EHRs systems in different scenarios. J Med Syst 2012;36(6):3777–82.
[43] BMWI Germany. GAIA-X: policy Rules and Architecture of Standards. Federal Ministry for Econ Affairs and Energy Germany 2020.
[44] Eggers G, Fondermann B, Maier B, Ottradovetz K, Pfrommer J, Reinhardt R, et al. GAIA-X: technical Architecture. Federal Ministry for Econ Affairs and Energy (BMWi) 2020.
[45] Fox E. Predominance of the curative model of medical care: a residual problem. JAMA 1997;278(9):761–3.
[46] Zheng Y, Ding X, Poon CCY, Lo BPL, Zhang H, Zhou X, et al. Unobtrusive sensing and wearable devices for health informatics. IEEE Trans Biomed Eng 2014;61(5):1538–54.
[47] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med 2020;3(1):1–10.
[48] Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc AMIA Symp 2007:26–30.
[49] Goddard K, Roudsari A, Wyatt JC. Automation bias - a hidden issue for clinical decision support system use. Stud Health Technol Inform 2011;164:17–22.
[50] Khalifa M, Zabani I. Improving utilization of clinical decision support systems by reducing alert fatigue: strategies and recommendations. Stud Health Technol Inform 2016;226:51–4.
[51] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30(7):1145–59.
[52] Thambawita V, Jha D, Hammer HL, Johansen HD, Johansen D, Halvorsen P, et al. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. ACM Trans Comput Healthc 2020;1(3). 17:1-17:29.
[53] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology, 13. MIS Q; 1989. p. 319–40.
[54] Bertl M. News analysis for the detection of cyber security issues in digital healthcare. Young Inf Sci 2019;4:1–15.
[55] Burke W., Oseni T., Jolfaei A., Gondal I. Cybersecurity indexes for eHealth. In: Proceedings of the Australasian Computer Science Week Multiconference [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2021 May 22]. p. 1–8. (ACSW 2019). Available from: 10.1145/3290688.3290721.
[56] Merhi MI. An evaluation of the critical success factors impacting artificial intelligence implementation. Int J Inf Manag 2022:102545.
[57] Vassigh K, Voracek D, Johnson M, Amato D, Frerking M, Beauchamp P, et al. Final report of the NASA technology readiness assessment (TRA) study team [Internet]. NASA 2015. May [cited 2021 May 19]. Available from, http://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level.
[58] TRL [Internet]. EURAXESS. 2020 [cited 2022 Nov 20]. Available from: https://euraxess.ec.europa.eu/career-development/researchers/manual-scientific-entrepreneurship/major-steps/trl.
[59] Bertl M., Shahin M., Ross P., Draheim D. Finding indicator diseases of psychiatric disorders in BigData using clustered association rule mining. In: Proceedings of ACM SAC Conference (SAC'23). Tallinn, Estonia: ACM; 2023.

Appendix 4

[IV]

M. Bertl, K. J. I. Kankainen, G. Piho, D. Draheim, and P. Ross. Evaluation of Data Quality in the Estonia National Health Information System for Digital Decision Support. In *Proceedings of the 3rd International Health Data Workshop*. CEUR-WS, 2023

Evaluation of Data Quality in the Estonian National Health Information System for Digital Decision Support

Markus Bertl[1,*,†], Kristian Juha Ismo Kankainen[1,†], Gunnar Piho[2], Dirk Draheim[2] and Peeter Ross[1,3]

[1]Department of Health Technologies, Tallinn University of Technology, Ehitajate tee 5, Tallinn, 12616, Estonia
[2]Department of Software Science, Tallinn University of Technology, Ehitajate tee 5, Tallinn, 12616, Estonia
[3]East Tallinn Central Hospital, Ravi 18, Tallinn, 10138, Estonia

Abstract

Following the implementation of Electronic Medical Records (EMR), the amount of digital health data has increased significantly in recent decades. This trend creates an opportunity to share data between different healthcare parties for primary and secondary use. However, the quality of this data is often questioned, and data reuse is still rare. This study evaluates the frequency of the use and quality of health data stored in the Estonian Health Information System (EHIS), which is one of the most advanced digital health platforms (DHP) in the world. We collected usage data of the EHIS from its initial release in 2008 till 2021. Comparing 2016 to 2021, the number of documents per year pushed into the EHIS has nearly doubled. But also approximately nine times more patients and five times more health professionals queried data from the EHIS. This increase in read access indicates that both groups find valuable information from the system. To investigate this further, data from patients with common diseases like stroke, cancer, or diabetes have been queried, analyzed, and compared against the actual data needs from the point of healthcare professionals and natural persons. Contradictory to the claim mentioned above, the manual analysis of the queried data sometimes showed poor data quality and missing information, especially discrepancies between the structured and unstructured parts of the documents shared through DHP. As an example of varying data quality, we looked at how smoking behavior is reported, both in structured form and in free text form in the queried data. We analyzed how the data quality of smoking behavior data shifts from document to document using the nine data quality dimensions of the Data Quality Vector. The data quality is shown to shift in 7 dimensions. While humans seem to be able to screen the data and resolve inconsistencies effectively, the data quality issues present make data reuse for tasks like AI training for digital decision support systems challenging.

Keywords

Data Quality, EHR (Electronic Health Record), Estonian National Health Information System, Digital Decision Support (DDSS), Artificial Intelligence (AI), Machine Learning (ML), Medical Data Reuse, Primary Use, Secondary Use

HEDA 2023: the 3rd International Workshop on Health Data (https://conf.researchr.org/home/staf-2023/heda-2023). Co-located with STAF 2023, 18–21 July, Leicester, United Kingdom.

*Corresponding author.

†These authors contributed equally.

markus.bertl@taltech.ee (M. Bertl); kristian.kankainen@taltech.ee (K. J. I. Kankainen); gunnar.piho@taltech.ee (G. Piho); dirk.draheim@taltech.ee (D. Draheim); peeter.ross@taltech.ee (P. Ross)

0000-0003-0644-8095 (M. Bertl); 0000-0002-0551-927X (K. J. I. Kankainen); 0000-0003-4488-3389 (G. Piho); 0000-0003-3376-7489 (D. Draheim); 0000-0003-1072-7249 (P. Ross)

1. Introduction

Estonia is a country in the north of Europe with 1.3 million citizens, approximately 4,500 physicians, and healthcare costs, which made up for about 7.5% of the annual GDP in 2021 [1]. The Estonian health system is based on mandatory, solidarity-based insurance and healthcare providers which operate under private law [2]. The Estonian digital health platform (DHP), called the Estonian nation-wide health information system (EHIS), has been operational since 2008 and allows secure and trusted online access to medical data, different kinds of medical documents, prescriptions, and medical images of virtually every Estonian resident from birth to death. It is fully integrated into the Estonian e-government systems, which provides a digital identity to every citizen, secure authentication methods, the possibility to link data according to the once-only principle, and other mature e-services [3, 4]. Instead of one large, centralized database, the EHIS comprises different federated and mutually independent systems. One is the nationwide electronic health record (EHR) system, which began the ongoing standardization of health-related data in Estonia [5]. In the central EHR, patient data is saved based on international standards like HL7 CDA[1], DICOM[2], LOINC[3], ICD-10[4] and SNOMED-CT[5]. The EHIS uses HL7 CDA as its data collection format. The CDA structure not only permits data capture in structured form but also allows to add medical data in unstructured free text format. Data sent to the EHIS is digitally signed or stamped by either the physician or the healthcare institution, which ensures accountability of the provided information. The data can be queried either directly from the data warehouse for statistical purposes and research, via API for eHealth applications like Hospital Information Systems (HIS), or via Web UIs like the patient portal over which residents of Estonia can view medical data from healthcare providers, referral letters, prescriptions, or fill out health declarations before an appointment. An overview of the first ten years of the EHIS can be found in [6].

As of today, the data collection process works as follows – The primary data sources for the EHIS are the electronic medical records (EMRs) of healthcare providers. Data is entered into the EMR by doctors and nurses or automatically transmitted from digital data sources such as laboratory equipment, etc. The data are entered in different modes: as free text, numeric data, including different codes (ICD-10, etc.), as graphs (ECG, etc.), or images (radiology, endoscopy, etc.). In order to share data with other institutions, the EMR exports data and digital documents in accordance with established standards (HL7 CDA, LOINC, etc.) for nationwide use and pushes them to the applications of various data consumers. One data consumer is the EHIS. Another data consumer is the Estonian Health Insurance Fund (EHIF), to which the ICD-10-coded diagnoses from the EMR are transmitted for billing purposes.

Digital decision support describes computer-based systems that bring together information from various sources, assist in the organization and analysis of information, and facilitate the evaluation of assumptions underlying the use of specific models [7]. Digital Decision Support Systems (DDSSs) could be divided by their goal of using them either as data capture aids or data

[1] http://www.hl7.org/
[2] https://www.dicomstandard.org/
[3] https://loinc.org/
[4] https://icd.who.int/browse10/
[5] https://www.snomed.org/

analysis and presentation tools. They can, for instance, be based on summarizing or visualizing data like the patient summary (Andmevaatur - data viewer in Estonian) functionality of the EHIS, or based on AI-based decision technology like rule-based expert systems [8], machine learning [9], or deep learning [10]. Regardless of the implementation flavor, data is needed for them to work accurately. One would expect that a sufficiently large amount of data to train and operate DDSSs is available through DHPs. Nevertheless, adoption rates of DDSSs are rather low [11]. AI algorithms for DDSS in healthcare itself, however, seem to perform sufficiently accurately [12, 13]. Besides having a holistic approach that includes domain experts from both the medical and the IT side, insufficient data quality has been found as one of the main barriers [11, 14]. Until now, there are two DDSSs operational in Estonia: the drug-drug interaction alert service (Inxbase[6]) and clinical decision support for primary healthcare physicians (EBMeDS[7]). Inxbase is part of the e-prescription services and uses manually defined rules to alert physicians if they prescribe medication that could interact with pharmaceuticals prescribed by other physicians [15]. There is currently no AI-based DDSS trained on data from the EHIS. Therefore, this research investigates the data quality of the EHIS in Estonia and assesses if the data stored there would even be usable for DDSSs.

2. Method

The EHIS has been chosen as the study object of this research because it is one of the most advanced nationwide DHPs in the world [16]. Therefore we assume it to be representative of the state-of-the-art in terms of data capture and data quality. We analyzed two parameters of the EHIS in this research:

- **Use** of saved health data measured by counting all queries made to the EHR from healthcare professionals through their EMRs and patients through the online accessible patient portal[8] of the EHIS. Queries can be, for instance, access to lab results, patient documentation, prescribed medication, or vaccination certificates.
- **Quality** of the captured data, especially to analyze the difference between structured and unstructured data, was measured by a Data Quality Vector (DQV). For this analysis, we decided to apply the DQV to the data of patients whose smoking status has been captured. Smoking is a highly relevant health factor and, in the EHIS case, can be documented both in free text in the EMR or in structured form in the health declaration of the EHR. We analyzed the entries of five randomly selected patients (12 documents in total) in this research to obtain preliminary results about the data quality.

The primary use of data is defined as data used directly for patient care and/or healthcare activities (including self-care). In contrast, secondary use (also called data reuse, multiple use, and further use) is defined as all data use that is not directly linked to patient care [17].

The Data Quality Vector (DQV) [18] offers a multi-dimensional view of data quality. Its nine data quality dimensions (Table 1) unify, according to its authors, all data quality dimensions

[6] https://www.medbase.fi/en/professionals/inxbase
[7] https://www.ebmeds.org/en/
[8] https://www.digilugu.ee/login?locale=en

Table 1
The nine dimensions of data quality according to the Data Quality Vector [18]

Dimension	Description
Completeness	The degree to which relevant data is recorded
Consistency	The degree to which data satisfies specified constraints and rules
Duplicity	The degree to which data contains duplicate registries representing the same entity
Correctness	The degree of accuracy and precision where data is represented with respect to its real-world state
Timeliness	The degree of temporal stability of the data
Spatial stability	The degree to which data is stable among different populations
Contextualization	The degree to which data is correctly/optimally annotated with the context in with it was acquired
Predictive value	The degree to which data contains proper information for specific decision-making purposes
Reliability	The degree of reputation of the stakeholders and institutions involved in the acquisition of data

proposed by other researchers previous to 2012. We used the DQV to assess in which dimensions data quality shifts occur between documents over time. Shifts were assessed between unstructured text and structured data, as well as between sequential documents.

The analyzed data concerns the smoking behavior of the subject of care, either in a structured form as part of health declarations or as free text excerpts as part of clinical reports (discharge summaries, referrals, etc.). The data is grouped by individual and includes all clinical documents about the person that was reported to the EHIS during the year 2019. The detailed inclusion criteria were: age 30–70 years, diagnosis of chronic disease (ICD-10 codes I00–I99, C00–C97, E10–E14). The initial sample size was 90 randomly selected individuals but evenly distributed across the diagnosis groups. The sample size was further decreased to 59 patients by filtering out only those with available data on smoking behavior. Data on smoking behavior was discovered by text search and annotated semantically by hand, otherwise as structured data in the health declaration form. The health declaration is a patient-reported questionnaire and is the basis for health certificates. Of the 59 patients with data on smoking behavior, only five had health declarations, whereas 57 had smoking behavior mentioned in free text. Three health declarations out of the five overlapped with information from free text. Of the two health declarations that provided smoking behavior without it being also mentioned in free text, one expressed smoking, and one expressed non-smoking status. We set up the DQV framework as follows. The analyzed documents were characterized as time-stamped and reported by different healthcare providers. Our analysis considers this time dependency, and although our data has been gathered in retrospect, we analyze it as if it was collected in a continuous data flow. To emulate decision support from the point of view of the document writer, we impose an imaginary constraint on whether the data has been available during the writing. This was judged by the look of the text, e.g., whether it is a copy-paste. The DQV was then used to analyze the documents in the following way. For each patient with data available on smoking behavior, the documents were ordered according to time. Thereafter the smoking behavior of each document was assessed and compared with the information mentioned in the succeeding document using each of the nine data quality dimensions.

The Estonian Human Research Ethics Committee (TAIEK) of the Institute for Health Development (Decision No. 1.1-12/186) approved the research design and data usage for this study.

3. Results

3.1. Use of Health Data

Figure 1 presents the number of documents added to the EHIS per year, the number of documents accessed by patients over the patient portal, and the number of queries from healthcare professionals from the initial launch of the EHIS in 2008 until 2021. While the number of documents pushed to the EHIS seems to reach a peak, the queries from both patients and doctors are still increasing sharply. The red line, which represents the number of queries from health professionals, only accounts for queries that have been actively and knowingly performed to get information from the EHR. The number does not contain system requests which are automatically performed during the clinical process. The blue line represents the number of queries performed through the patient portal of the EHIS, so it shows how much EHR data the patients view. It is worth mentioning that the number of queries is not equal to the number of natural persons logged into the patient portal, as several queries are usually made during a single patient portal session. Also, it is important to consider the COVID-19 pandemic when interpreting the numbers in Fig. 1. Vaccination certificates and lab test results are also part of the EHIS. Accessing them also contributed to the rise in patient queries in 2020 and 2021.

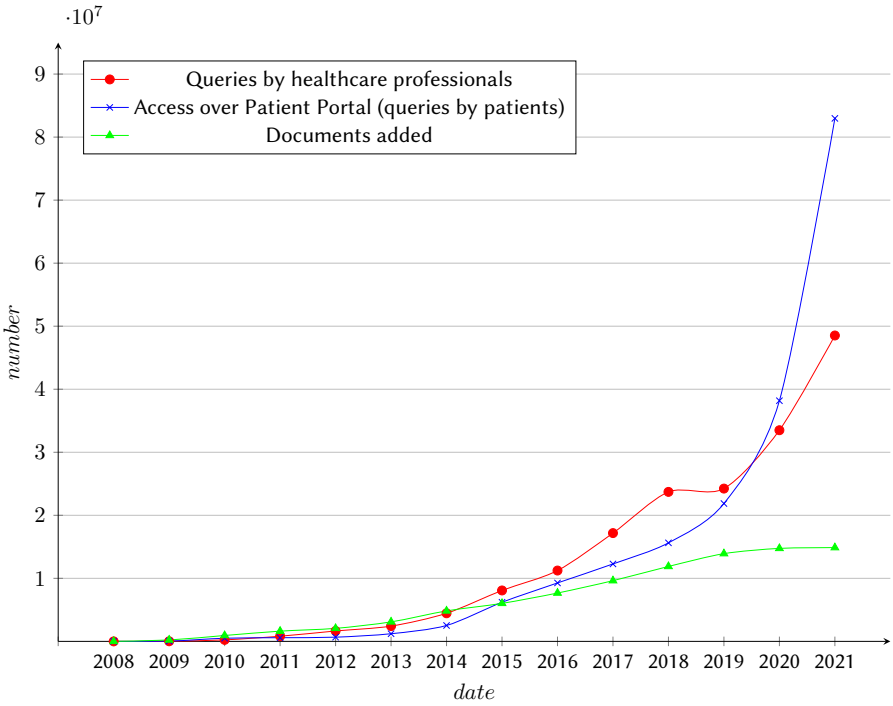


Figure 1: Data access of Estonian National Health Information System 2008-2021

3.2. Shifts in data quality occurring between documents

In the following subsections, we report on five illustrative findings of how data quality was assessed to shift between the analyzed documents according to the nine dimensions of the Data Quality Vector introduced above.

3.2.1. Shift between discharge summary and structured health declaration – the case of not smoking

In this example, we analyze two documents. The first document is the unstructured text of an anamnesis section of a discharge summary. The text states in one sentence both facts of non-smoking and alcohol-drinking behavior. The second document is a structured health declaration form that was filled in four months later. The structure of the health declaration form is such that data on smoking behavior and drinking behavior are in separate fields. Shifts in the following data quality dimensions can be observed (also Table 2):

Completeness does not shift, as the smoking status "not smoking" is semantically fully intact on both documents.
Consistency shifts in a technical sense as the first document is not machine-readable while the second is.
Duplicity of data does not occur, as smoking status changes over time.
Correctness of the data does not shift.
Timeliness does not apply, as smoking status changes over time.
Spatial stability shifts similar to the consistency dimension (text VS structured).
Contextualization of data does shift because the tight connection to alcohol-drinking behavior is not preserved in the health declaration.
Predictive value of the data does not shift. This dimension does not apply as nothing predicts a change in smoking behavior.
Reliability of the data can be said to shift, as health declarations are often filled by the patient.

Table 2
Shifts in data quality dimensions between discharge summary and structured health declaration – the case of not smoking (example 1, section 3.2.1).

Dimension	Shift
Completeness	No
Consistency	Yes
Duplicity	No
Correctness	No
Timeliness	-
Spatial stability	Yes
Contextualization	Yes
Predictive value	-
Reliability	Yes

3.2.2. Shift between discharge summary and structured health declaration – the case of smoking

In this example, we again have an anamnesis and a health declaration form filled in four months later. The first document states in anamnesis vitae the smoking longevity (“long-term

smoker"), temporal length ("30 years"), and the smoking amount ("one pack per day"). The second document is a health declaration form filled in four months later stating the patient is a smoker, the length in years ("30"), and the smoking amount in cigarettes per day ("2"). Shifts in the following data quality dimensions can be observed (also Table 3):

Completeness was unaffected, as all three semantic attributes (smoking status, length, and amount) remained intact. It could be argued that some interpretation of longevity (explicitly marked "long-term") is lost, although the length in years stays the same.

Consistency shifts as different units are used for the smoking amount (packs vs. cigarettes). Another shift could be argued with the loss of qualifier ("long-term").

Duplicity does not apply, as smoking status changes over time.

Correctness of the data can be analyzed both ways. If the smoking behavior has not changed, the number of cigarettes is a typo, as one pack (in Estonia) equals 20 cigarettes. In case of a change in smoking behavior, a decrease has occurred from 20 to 2 cigarettes per day.

Timeliness does not apply.

Spatial stability shifts similarly to consistency (text VS structured).

Contextualization did not change as both contexts can be interpreted as general knowledge about the patient's lifestyle.

Predictive value of the data does not shift. This dimension does not apply as nothing predicts a change in smoking behavior.

Reliability of the data can be said to shift, as health declarations are often filled by the patient.

Table 3
Shifts in data quality dimensions between discharge summary and structured health declaration – the case of smoking (example 2, section 3.2.2).

Dimension	Shift
Completeness	No
Consistency	Yes
Duplicity	No
Correctness	No/Yes
Timeliness	-
Spatial stability	Yes
Contextualization	No
Predictive value	-
Reliability	Yes

3.2.3. Shift between two inpatient discharge summaries

The first document states in the treatment synopsis of an inpatient rheumatology discharge summary a recommendation to stop smoking, among other recommendations. The second document is an inpatient cardiology discharge summary and states twice in the anamnesis the fact of being a smoker (first in the problem list and then in a separate smoking status field). Shifts in the following data quality dimensions can be observed (also Table 4):

Completeness was affected as the first document contained only the cessation recommendation, and the second document stated only the fact of being a smoker.

Consistency was breached as our hierarchy rules state smoking status should come before cessation recommendation.

Duplicity was found inside the second document without data reuse being evident; rather, the information was presented in two contextualizations.

Correctness of the data can be both ways: it can be seen as dependent on reasoning capabilities: if cessation recommendation implies being a smoker, then all is correct. Another view would be to allow non-linearity in that cessation recommendations can correctly be given to anyone, also non-smokers.

Timeliness was not evident in the data.

Spatial stability was analyzed similarly to correctness – it relies on reasoning capabilities, e.g., the rules specified by the consistency dimension.

Contextualization was different for each occurrence: implicitly in cessation recommendation (treatment), explicitly in the problem list, and separately as smoking status.

Predictive value of the data does not shift. This dimension does not apply as nothing predicts a change in smoking behavior.

Reliability of the data does not shift.

Table 4
Shifts in data quality dimensions Shift between two inpatient discharge summaries (in example 3, section 3.2.3).

Dimension	Shift
Completeness	Yes
Consistency	Yes
Duplicity	Yes
Correctness	-
Timeliness	-
Spatial stability	Yes
Contextualization	Yes
Predictive value	-
Reliability	-

3.2.4. Richness of the contextualization dimension

This example consists of only one document; therefore, no shift can be analyzed. Instead, our intention here is to highlight the value and richness of contextuality from the healthcare professional's point of view. We found in one document that the smoking behavior was stated in a comment next to the structured data fields with elevated blood pressure and pulse. The textual comment had a nomenclature code for a cardiovascular observable and an interpretation code for normal. The free text of the comment stated the patient not being a smoker.

3.2.5. Shifts occurring between multiple documents

In this example, we could trace smoking behavior across five different documents.

- The first document is the anamnesis section of a referral that states the longevity of smoking ("long-term smoker").
- The second document is the anamnesis section of an outpatient visit. It duplicates exactly the text from the referral and adds no more information. This we analyze as a reuse of timely available data.

- The third document is the anamnesis section of an inpatient discharge summary stating smoking status, adding an approximate numerical quantification of longevity ("more than 20 years") and the amount in packs per day as a span ("1–1.5").
- The fourth document is the anamnesis section of a pulmonology outpatient discharge summary. It duplicates the exact phrase from the previous document but adds to it the current trend of smoking amount ("has tried to cut down lately").
- The fifth document is a general practitioner outpatient discharge summary treatment regimen. It does not mention smoking status but states the importance that the patient stops smoking, e.g., is an instruction on smoking cessation.

Refer to Table 5 for our analysis of the shifts from document to document according to the DQV dimensions.

Table 5
Shifts in data quality dimensions between multiple documents (example 5, section 3.2.5).

Dimension	Shift I–II	Shift II–III	Shift III–IV	Shift IV–V	Shift V-VI
Completeness		No shift	Adds data	Adds data	
Consistency	Is consistent	No shift	Shift to more precise granularity	No shift	
Duplicity		Duplicates		Duplicates	
Correctness					
Timeliness		Was timely		Was timely	
Spatial stability					
Contextualization	Anamnesis in referral	Anamnesis in referral	Anamnesis in inpatient discharge summary	Anamnesis in outpatient discharge summary	Needed self-care activity, Treatment regime in GP discharge summary
Predictive value	-	-	-	-	-
Reliability	-	-	-	-	-

4. Discussion

Our data quality vector analyses show clearly that data quality shifts in several dimensions between documents. We observed shifts in the following dimensions: Completeness, Consistency, Duplicity, Correctness, Spatial stability, Contextualization, and Predictive value. It is evident that the granularity of information changes throughout the clinical process. If, for example, only the smoking status is needed for a decision, it is found in more documents than the more precise knowledge of how many cigarettes the patient smokes daily. Additionally, having information in both structured and unstructured forms creates redundancies, leading to inconsistency. This not only introduces challenges to data usage for decision support but also makes secondary use difficult because the inconsistently structured or even unstructured data is hard to aggregate (e.g., querying the average number of smoked cigarettes per age group).

The number of new documents added per year to the EHR system seems to be reaching its current maximum, with no sharp increases observed since 2019. In contrast, a sharp increase in

the number of queries can be observed for the same time span. This usage pattern indicates to us that both patients and healthcare professionals have been getting useful information from the system in recent years. Otherwise, the users would not query it increasingly. If the system did not bring a benefit, people would use it less and query rates should stagnate or even decrease. Such a trend is visible from 2008 to approximately 2014 in Figure 1. The EHIS was just launched back then and did not contain enough useful information for patients and healthcare professionals to yield high query counts. The trend of rising query rates, combined with our findings of inconsistent and potentially low-quality data, raises the question of why medical professionals still use the EHR system increasingly. It is a clinical routine that healthcare professionals have to use as many data sources as reasonably possible about the patient's health status to make medically relevant decisions. So far, medical professionals' education has emphasized the importance of reading previous patient files and test results. This means that in the Estonian case, healthcare professionals are approaching the patient data mainly in a conventional manner, not benefiting from the full spectrum of digital data-sharing opportunities. However, the latter (e.g., the use of DDSS in the clinical process) is possible only if the collected data is standardized and structured, making it available for computer processing. Also, it could be argued that humans can make better sense of the available low-quality textual data by intuitively determining which interpretation of the inconsistent data is most likely correct – a problem that is still hard for AI algorithms. Recent advantages in deep learning, especially deep natural language processing (NLP), do allow the use of unstructured data. However, whether those methods give the needed accuracy is still questionable. Using NLP methods to structure unstructured health data by extraction can also introduce additional inconsistencies and shifts in other data quality dimensions. As one example, the indicated practice of exchanging (unstructured) data within referral documents (see 3.2.5), where each receiving healthcare professional adds more detail and sends the elaborated data with a new referral. This practice leads to a data integration situation that is very hard to coordinate: the previously known data is duplicated, and the new data elements are rooted within their own contexts creating instability in both the spatial and timeliness dimensions. It is not the structuring of data that is hard, but instead, the coordination and interpretation. The shift in spatial stability leads to the question of which source should be accounted for, and the shift in the timeliness dimension leads to the question of when to account for what data. These shifts, in turn, affect the completeness and consistency dimensions. Therefore, using NLP technologies to structure textual health data would introduce additional risks for a DDSS since it potentially would work on tainted data.

One of the few pieces of information available in a quality-controlled format is demographic information linked from other e-government registered and the mandatory ICD-10-coded diagnosis, which needs to be recorded for billing purposes at each physician visit. The challenge for DDSSs is that much of the more granular information in the EHIS is still stored in a free text format instead of a machine-readable, structured form. Humans can interpret these textual descriptions, but they are not machine-understandable. This makes data reuse challenging. Structured data would be desirable to train AI algorithms for decision support, effectively query data, or perform statistical analyses. If we assume some mechanism that would make the free text of clinical documents machine-understandable, then in the case of smoking behavior, our results show the need not only to reuse but also to manipulate the data by later refining the semantics (pt is smoker > pt smokes for X years > pt smokes X cig/day > pt has cut down the

amount of cig/day > pt stopped smoking > pt has not smoked for X months). Our analysis supports the hypothesis that one document is not always enough for a granular understanding of smoking status, but rather a cumulative view should exist. But data aggregation presupposes structured data instead of free text. To maximize the usefulness of decision support, not only one axis, like smoking status (yes/no), needs to be queryable through structured data, but also at least a second, time-based dimension containing more granular data like the number of cigarettes/day. This would introduce more information for AI-based algorithms to train on and potentially allow more accurate predictions.

Generalizing this understanding to other health data, the rise in patient's document retrievals might also be due to difficulty finding the right information, resulting in multiple searches for the document containing the needed information. For example, more documents contain information answering yes/no questions, whereas few documents contain more granular information.

We want to highlight that this research only presents a preliminary analysis that probably does not cover all data quality issues in the current DHP. Our main goal was to show that there are severe data quality issues even in those small, random samples. Based on the methodology described, a more detailed data quality analysis of a larger cohort of patients will follow.

5. Conclusion

Our analysis shows that the use of nationwide electronic health records embedded in a digital health platform is well accepted and widely used by healthcare professionals and patients, despite the sometimes questionable quality of the data. The number of queries to the EHIS is rising, which shows increased use and indicates that people are finding helpful information. We discovered shifts in seven of nine data quality dimensions by analyzing individual documents in detail. The shifts express, among other, information being added upon and made more precise, and inconsistencies between the structured and unstructured (free text) parts of an entry to the EHIS. Humans can make sense of the shifting data quality and unstructured data by using abductive reasoning (intuitively using their knowledge to find the most likely interpretation of the available information). This is challenging for machines, making the data difficult for tasks like AI training, effectively querying data, or performing statistical analyses. For this, high-quality, structured data would be needed. Although specific mandatory structured data fields in the EHR, like ICD-10 coded diagnosis, can be utilized for DDSSs, structured data on more complex information is often still not available.

Acknowledgments

The Estonian Human Research Ethics Committee (TAIEK) of the Institute for Health Development (Decision No. 1.1-12/186) approved the research design and data usage for this study.

This work in the project 'ICT programme' was supported by the European Union through the European Social Fund and the Norway Grants Program "Green ICT" (Nmb. F21009).

References

[1] T. Habicht, K. Kahur, K. Kasekamp, K. Köhler, M. Reinap, A. Võrk, R. Sikkut, L. Aaben, E. Van Ginneken, E. Webb, et al., Estonia: health system summary, 2022 (2023).

[2] T. Lai, T. Habicht, M. Jesse, Monitoring and evaluating progress towards universal health coverage in estonia, PLoS Medicine 11 (2014) e1001677.

[3] R. Krimmer, T. Kalvet, M. Toots, A. Cepilovs, E. Tambouris, Exploring and demonstrating the once-only principle: a european perspective, in: Proceedings of the 18th annual international conference on digital government research, 2017, pp. 546–551.

[4] S. Lips, V. Tsap, N. Bharosa, R. Krimmer, T. Tammet, D. Draheim, Management of national eID infrastructure as a state-critical asset and public-private partnership: Learning from the case of Estonia, Information System Frontiers (2023). doi:10.1007/s10796-022-10363-5.

[5] M. Tiik, P. Ross, Patient opportunities in the estonian electronic health record system, in: Medical and Care Compunetics 6, IOS Press, 2010, pp. 171–177.

[6] J. Metsallik, P. Ross, D. Draheim, G. Piho, Ten years of the e-health system in estonia, in: A. Rutle, Y. Lamo, W. MacCaull, L. Iovino (Eds.), CEUR Workshop Proceedings, volume 2336, 3rd International Workshop on (Meta)Modelling for Healthcare Systems (MMHS), 2018, pp. 6–15. URL: ceur-ws.org/Vol-2336/MMHS2018_invited.pdf.

[7] V. Sauter, Decision support systems: an applied managerial approach, John Wiley & Sons, Inc., 1997.

[8] M. Bertl, M. Shahin, P. Ross, D. Draheim, Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining, in: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 826–833. URL: https://doi.org/10.1145/3555776.3577594. doi:10.1145/3555776.3577594.

[9] M. Bertl, P. Ross, D. Draheim, Predicting psychiatric diseases using autoai: A performance analysis based on health insurance billing data, in: Database and Expert Systems Applications, Springer International Publishing, Cham, 2021, pp. 104–111.

[10] M. Bertl, N. Bignoumba, P. Ross, S. B. Yahia, D. Draheim, Evaluation of deep learning-based depression detection using medical claims data, SSRN (2023).

[11] M. Bertl, P. Ross, D. Draheim, Systematic ai support for decision making in the healthcare sector: Obstacles and success factors, Health Policy and Technology (2023). doi:https://doi.org/10.1016/j.hlpt.2023.100748.

[12] M. Bertl, J. Metsallik, P. Ross, A systematic literature review of ai-based digital decision support systems for post-traumatic stress disorder, Frontiers in Psychiatry 13 (2022). URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.923613. doi:10.3389/fpsyt.2022.923613.

[13] M. Bertl, P. Ross, D. Draheim, A survey on ai and decision support systems in psychiatry – uncovering a dilemma, Expert Systems with Applications 202 (2022) 117464. URL: https://www.sciencedirect.com/science/article/pii/S0957417422007965. doi:https://doi.org/10.1016/j.eswa.2022.117464.

[14] M. Bertl, T. Klementi, G. Piho, P. Ross, D. Draheim, How domain engineering can help to raise decision support system adoption rates in healthcare, 2023.

[15] K. Kõnd, A. Lilleväli, et al., E-prescription success in estonia: The journey from paper to phamacogenomics, Eurohealth 25 (2019) 18–20.

[16] F. Colombo, J. Oderkirk, L. Slawomirski, Health information systems, electronic medical records, and big data in global healthcare: Progress and challenges in oecd countries, Handbook of global health (2020) 1–31.

[17] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, C. U. Lehmann, Clinical data reuse or secondary use: current status and potential future progress, Yearbook of medical informatics 26 (2017) 38–52.

[18] C. Sáez, J. Martínez-Miranda, M. Robles, J. M. García-Gómez, Organizing Data Quality Assessment of Shifting Biomedical Data, Quality of Life through Quality of Information (2012) 721–725. doi:10.3233/978-1-61499-101-4-721, publisher: IOS Press.

Appendix 5

[V]

M. Bertl, P. Ross, and D. Draheim. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *Database and Expert Systems Applications*, pages 104–111. Springer International Publishing, 2021

Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data

Markus Bertl[1( )], Peeter Ross[1], and Dirk Draheim[2]

[1] Department of Health Technologies, Tallinn University of Technology, Tallinn, Estonia
{mbertl,peeter.ross}@taltech.ee
[2] Information Systems Group, Tallinn University of Technology, Tallinn, Estonia
dirk.draheim@taltech.ee

Abstract. Digital transformation enables a vast growth of health data. Because of that, scholars and professionals considered AI to enhance quality of care significantly. Machine learning (ML) algorithms for improvement have been studied extensively, but automatic artificial intelligence (autoAI/autoML) has been widely neglected. AutoAI aims to automate the complete AI lifecycle to save data scientists from doing low-level coding tasks. Additionally, autoAI has the potential to democratize AI by empowering non-IT users to build AI algorithms. In this paper, we analyze the suitability of autoAI for mental health screening to detect psychiatric diseases. A sooner diagnosis can lead to cost savings for healthcare systems and decrease patients' suffering. We evaluate AutoAI using the open-source machine learning library auto-sklearn, as well as the commercial Watson Studio's AutoAI platform to predict depression, post-traumatic stress disorder, and psychiatric disorders in general. We use health insurance billing data from 83,986 patients with a total of 687,697 ICD-10 coded diseases. The results of our research are as follows: (i) on average, an accuracy of 0.6 (F_1–score 0.58) with a precision of 0.61 and recall of 0.56 was achieved using auto-sklearn. (ii) The evaluation metrics for Watson Studio's autoAI were 0.59 accuracy, 0.57 F_1–score, a precision of 0.6, and a recall of 0.55. We conclude that the prediction quality of autoAI in psychiatry still lacks behind traditional ML approaches by about 24% and is therefore not ready for production use yet.

Keywords: Artificial intelligence · AI · Machine learning · ML · AutoAI · AutoML · IBM Watson AutoAI · Auto-sklearn · Decision support systems · Psychiatry · Depression · Post-traumatic stress disorder (PTSD)

1 Introduction

Artificial Intelligence (AI) is nowadays a major driver for innovation in digital government and will play a significant role in tackling the challenges our society

C. Strauss et al. (Eds.): DEXA 2021, LNCS 12923, pp. 104–111, 2021.
https://doi.org/10.1007/978-3-030-86472-9_9

is currently facing. This especially applies to the healthcare sector. The creation of health data is rising year by year. Together with the drastic increase in computing power and the rising acceptance of AI by the general public, research about AI-driven digital decision support systems (DDSS) gets more and more popular. The Cambridge Dictionary defines AI as "the study of how to produce computers that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, solve problems, and learn" [14]. Sauter 1997 defines DDSSs as "computer-based systems that bring together information from a variety of sources, assist in the organization and analysis of information and facilitate the evaluation of assumptions underlying the use of specific models" [15]. Especially in mental health, this has enormous potential to optimize patient care. The prevalence of mental illnesses, suffering, and stigmatization are high – at the same time, diagnostic accuracy is low. 52.7% of people with depression are not correctly diagnosed in a primary care [11]. Mental illness has a severe impact on the society as a whole; In [8], Greenberg et al. estimate that major depressive disorder alone results in an economic burden of approximately $210.5 billion annually.

These obstacles, together with the exponential data growth, create increasing opportunities for DDSS. Recent meta-reviews showed that current research of DDSS promises high accuracy scores using ML algorithms [1,2]. The used data was mostly transformed and algorithms were specifically selected and tuned. Building, maintaining, and operating AI algorithms that way not only requires advanced data science skills but is also time-intensive.

The relatively new research area of automatic AI (autoAI), sometimes also called automatic machine learning (autoML), tackles this problem by providing systems that automate the whole ML lifecycle end to end (e.g., data preparation, feature engineering, model selection, pipeline optimization, and hyperparameter optimization). There are both open-source products (such as auto-sklearn [6], or autokeras [10]), as well as commercial products (such as IBM Watson Studio's AutoAI [9]) available. AutoAI speeds up the process of developing ML models and will be inevitable in the future of data science [17]. Additionally, autoAI can democratize AI by enabling non-technical users to apply AI technology easily because they do not need to understand the statistical background for selecting and tuning the right AI algorithm. AutoAI takes data, automatically transforms it as needed, selects a proper algorithm, and automatically tunes the hyperparameter of the algorithm. Currently, autoAI has not found wide adoption in scientific literature, nor clinical practice [18].

The described benefits led us to our research question on how autoAI algorithms in the healthcare sector perform compared to traditional approaches, and if autoAI is an alternative for increasing DDSS adoption rates by enabling a faster implementation. Our main contributions in this paper are as follows:

- A novel performance evaluation of two popular autoAI frameworks (open-source, as well as commercial) against a large, real-world dataset in psychiatry.
- We put this evaluation into the context of traditional, manual, machine learning approaches to test the suitability of autoAI for AI-based DDSSs.

The paper is organized as follows. In Sect. 2, we provide an overview of the used data set and describe the investigated autoAI frameworks and the development environment. In Sect. 3, we present the results of our evaluation. In Sect. 4, we discuss our approach and put the results in context of other research. Finally, we finish the paper with a conclusion in Sect. 5.

2 Method

2.1 Data

The data used was collected from the Estonian Health Insurance Fund. Our dataset includes information on sex, birth year, diagnosis, and diagnoses year and month from 83,986 adults (18 years or above) with a total of 687,697 diagnoses in 2019. The data consists of all publicly insured people in Estonia with a depression diagnosis, either single episode (F32), or recurrent (F33), an equally-sized random sample of people with other psychiatric disorders and no depression diagnosis, and an also equally sized random sample of people without any psychiatric diagnosis. Records with only one diagnosis were excluded.

Patients with and without the disease to predict were equally sampled, and test and training data were divided in a 1:4 ratio. Since our goal is a benchmark of how well autoAI performs on its own, we did no further data preparation or feature engineering on the training data. The test data was selected using proportionate stratified random sampling with the two strata 'healthy' and 'the disease to predict' in a 1:1 ratio, to decrease the risk of sampling bias. Additionally, this solves the problem of misleading accuracy values because of the oversampled 'healthy' group.

The diagnoses were coded using the International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) which is an international standardized coding system for reporting diseases [19]. Each ICD-10 code is structured based on an alpha character and two digits describing the category of the disease followed by a dot and further digits representing more details such as cause, location, severity, or other clinical information. As one example, F32.2 codes for major depressive disorder, single episode, severe without psychotic features. F stands for mental and behavioral disorders, F30–F39 code mood [affective] disorders where F32 is the sub-item for major depressive disorder, single episode. The '.2' in the end specifies the severity.

Our performance benchmark is based on the prediction of F32 (major depressive disorder, single episode), F33 (major depressive disorder, recurrent), F43 (reaction to severe stress, and adjustment disorders), and if any F-diagnosis present is present. For the prediction of each diagnosis we report:

$$precision = \frac{true\ positives}{true\ positives\ +\ false\ positives} \tag{1}$$

$$recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \tag{2}$$

$$F_1\text{-}score = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives} \tag{4}$$

Definitions of the measurements can be found in [16].

2.2 AutoAI Frameworks, Development Environment, and Configuration

Our development environment consisted of *JupyterLab 3.0.12* with *Python 3*, *Pandas 1.2.3*, and *auto-sklearn 0.12.4* with the Auto-Sklearn 2.0 classifier [5]. To analyze the models produced by auto-sklearn, we are using the Python package *PipelineProfiler* [13].

Auto-Sklearn was chosen because it is a state-of-the-art open-source autoAI framework claiming to outperform its competitors [6]. Auto-Sklearn works based on Bayesian optimization, meta-learning, and ensemble construction [6]. Auto-sklearn supports 15 classifiers, 14 feature pre-processing techniques, and 4 data pre-processing methods, with 110 hyperparameters [5,6]. We used a seed value of 7 for the Auto-Sklearn 2 classifier. Since auto-sklearn is non-deterministic, we execute each training five times to see how the different runs compare.

We also included the results from IBM Watson Studio's AutoAI, the autoAI platform on the IBM public cloud [9] to see if commercial platforms perform differently. Watson Studio's AutoAI supports 7 classifiers and 20 data transformations [9]. It uses a model-based, derivative-free global search algorithm, called RBfOpt [4] for hyperparameter optimization, in contrast to Auto-sklearn's Bayesian optimization. We configured accuracy as optimization metric for both frameworks. For both frameworks, no restrictions concerning time or memory were given. We ensured that enough RAM and disk space are available for each training run to finish without out-of-memory errors.

3 Results

3.1 Watson AutoAI

Table 1 gives an overview of the different classifiers' metrics resulting from Watson Studio's AutoAI on IBM Cloud. Since the F diagnoses data set was not supported by autoAI because of file size limitations, it is omitted in the results table. The average accuracy of all runs for all diseases is 0.59, with a F_1–score of 0.57, a precision of 0.6 and a recall of 0.55. In all cases, Watson Studio's AutoAI chose LGBM classifier (gradient boosting with leaf-wise tree-based learning) with first applying principal component analysis to the data.

Table 1. Watson AutoAI – performance.

Disease	Precision	Recall	F_1-score	Accuracy	Test data
F32	0.65	0.52	0.58	0.62	4331
F33	0.60	0.59	0.59	0.60	4325
F43	0.55	0.55	0.55	0.55	1195

3.2 Auto-Sklearn

Table 2 shows the aggregated metrics of the different classifier ensembles created by the five runs of auto-sklearn.

The average accuracy of all runs for all diseases is 0.6, 95% CI [0.596, 0.604], with a F_1–score of 0.58, a precision of 0.61 and a recall of 0.56. In the following subsections, we report more details of the construction and the metrics for each classifier ensemble. Since auto-sklearn is not deterministic, different runs can lead to different results. However, the average standard deviation of the accuracy values was with $\sigma = 0.0057$ small.

Table 2. Auto-sklearn classifiers – average performance.

Disease	Precision	Recall	F_1-score	Accuracy	Test data
F32	0.60	0.56	0.58	0.59	4331
F33	0.63	0.57	0.6	0.61	4325
F43	0.59	0.63	0.61	0.59	1195
F	0.63	0.47	0.53	0.60	17194

F32 - Major Depressive Disorder, Single Episode. During our five runs, auto-sklearn analyzed between 200 and 213 target algorithms for this classification task. On average, an accuracy of $\mu = 0.59$ with a standard deviation of $\sigma = 0.0055$ was achieved with five independent runs. The final ensemble consisted of 35 pipelines. Categorical and numerical transformers were used for pre-processing, Gradient Boosting [12], Random Forest [3], and Extra Trees [7] as classifiers.

F33 - Major Depressive Disorder, Recurrent. Auto-sklearn analyzed between 215 and 251 target algorithms for this classification task. On average, an accuracy of $\mu = 0.612$ with a standard deviation of $\sigma = 0.0084$ was achieved with five independent runs. The final ensemble consisted of 30 pipelines. Categorical and numerical transformers were used for pre-processing and Gradient Boosting [12] and Extra Trees [7] as classifiers.

F43 - Reaction to Severe Stress, and Adjustment Disorders. Auto-sklearn analyzed between 99 and 140 target algorithms for this classification task. On average, an accuracy of $\mu = 0.594$ with a standard deviation of $\sigma = 0.0089$ was achieved with five independent runs.

The final ensemble consisted of 18 pipelines. Categorical and numerical transformers were used for pre-processing, and Random Forest [3] and Gradient Boosting [12] as classifiers.

All F − Diagnoses. Auto-sklearn analyzed between 148 and 150 target algorithms for this classification task. On average, an accuracy of $\mu = 0.6$ with a standard deviation of $\sigma = 0.0$ was achieved with five independent runs. The final ensemble consisted of 27 pipelines. Categorical and numerical transformers were used for pre-processing and Gradient Boosting [12] as classifier.

4 Discussion

We demonstrated that auto-sklearn can be used to predict psychiatric diseases using health insurance billing data. However, the resulting evaluation metrics are behind the ones that are reported using traditional AI approaches. While we achieved an accuracy of 0.6 with auto-sklearn and 0.59 with Watson Studio's AutoAI Experiment, a recent literature survey reports an average accuracy of 0.84 for predicting psychiatric diseases [2]. Notable is, that the studies found by the named survey have a by far smaller sample size (mean of $\mu = 5569$ records with a standard deviation of $\sigma = 19194.28$ and a median of $\eta = 237$) than we used in our evaluation (687,697 records). While most research just has samples from one hospital or healthcare provider, we were able to use data from all patients with depressions in Estonia to test autoAI on a realistic, BigData sample from a whole country.

We observed that auto-sklearn is easy to use, even for non-IT professionals, and gives fast results without the need for data science skills. Watson Studio's AutoAI Experiment does not even require coding skills. All tasks can be carried out through a web interface on the IBM Cloud. Concerning the initial research question, our experiment showed that the tested autoAI frameworks still lag behind traditional approaches where data scientists develop and tune ML models. AutoAI can by no means replace data scientists. While autoAI offers functionality for feature engineering, model selection and tuning, data scientists still need to attend to the human side of AI model implementation like finding the right business problem to solve, analyze the requirements, and determining the superiority of an AI solution compared to currently used solutions.

Nevertheless, autoAI holds the potential to save data scientists time by automating basic, low-level tasks, enabling the professionals to focus on understanding the business problem and later on, the individual fine-tuning of the outputted autoAI models. This can decrease the time for data scientists to design new AI models. Since autoAI is a relatively new field in computer science, it is not fully matured and we expect it to deliver better results in the near future.

One limitation of our research is that we compare the autoAI results to studies using other data for evaluation. In further research, we will use the applied dataset for traditional machine learning algorithms, as well as deep learning algorithms to see how they perform compared to autoAI. Additionally, our research is limited to the binary classification functionality of auto-sklearn and IBM Watson Studio's AutoAI. Other libraries might perform differently. Because of that, our results are not generalizable for the whole autoAI area. Neither auto-sklearn nor Watson Studio's AutoAI support deep learning. Considering the complexity and high dimensionality of the dataset, deep learning algorithms could lead to better results. Another limitation comes from the way auto-sklearns optimization works. The system is non-deterministic, so that results may change between different runs. To get a comprehensive picture, further research needs to be done to compare these libraries and traditional AI approaches based on a standard benchmark dataset.

5 Conclusion

AutoAI is an emerging research field in computer science with high potential. The ease of use enables a faster development of AI algorithms without extensive data science or programming knowledge. Therefore, it contributes to democratizing AI-based solutions. One possible area where autoAI could be applied in the healthcare sector are DDSSs. Despite the theoretical potential, there is currently no thorough evaluation of autoAI on healthcare datasets.

We presented a novel evaluation of autoAI libraries based on real-world data. In our setting, the accuracy of autoAI (namely auto-sklearn classifier 2.0 and Watson Studio's AutoAI Experiment on IBM Cloud) without any human intervention could not achieve as good evaluation metrics as traditional approaches. Our main finding is that AutoAI's accuracy was, on average, 24% behind conventional techniques. Because of that, we argue that autoAI is not yet ready to be used for the detection of psychiatric diseases based on health insurance billing data. Manual tuning and especially domain knowledge is still needed to create an accurate machine learning model. Because of the ease of use, autoAI can serve as a rapid prototyping tool for ML. It gives an initial direction and can be seen therefore as a smart assistant for data scientists, saving their time for the human side of machine learning projects and advanced fine-tuning tasks after the initial model creation.

References

1. Bertl, M., Metsallik, J., Ross, P.: Digital decision support systems for post-traumatic stress disorder - implementing a novel framework for decision support systems based on a technology-focused, systematic literature review (2021). https://doi.org/10.13140/RG.2.2.12571.28965/1
2. Bertl, M., Ross, P., Draheim, D.: A survey on AI and decision support systems in psychiatry - uncovering a dilemma (2021). https://doi.org/10.13140/RG.2.2.10810.82880/2

3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Costa, A., Nannicini, G.: RBFOpt: an open-source library for black-box optimization with costly function evaluations. Math. Program. Comput. **10**(4), 597–629 (2018)
5. Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., Hutter, F.: Auto-sklearn 2.0: the next generation. arXiv:2007.04074 [cs, stat] (2020)
6. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Proceedings of NIPS 2015 - The 28th Annual Conference on Neural Information Processing Systems, pp. 1–9 (2015)
7. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
8. Greenberg, P.E., Fournier, A.A., Sisitsky, T., Pike, C.T., Kessler, R.C.: The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J. Clin. Psychiatry **76**(2), 155–162 (2015)
9. IBM: AutoAI-implementation details - IBM Watson studio (2021). https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai-details.html?audience=wdp
10. Jin, H., Song, Q., Hu, X.: Auto-keras: an efficient neural architecture search system. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1946–1956. ACM (2019)
11. Mitchell, A.J., Vaze, A., Rao, S.: Clinical diagnosis of depression in primary care: a meta-analysis. Lancet **374**(9690), 609–619 (2009)
12. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Front. Neurorobot. **7**, 21 (2013)
13. Ono, J.P., Castelo, S., Lopez, R., Bertini, E., Freire, J., Silva, C.: PipelineProfiler: a visual analytics tool for the exploration of AutoML pipelines. arXiv:2005.00160 [cs] (2020). http://arxiv.org/abs/2005.00160
14. Procter, P. (ed.): Cambridge International Dictionary of English. Cambridge University Press, Cambridge (1995)
15. Sauter, V.L.: Decision Support Systems for Business Intelligence. Wiley, Hoboken (1997)
16. Tohka, J., van Gils, M.: Evaluation of machine learning algorithms for health and wellness applications: a tutorial. Comput. Biol. Med. **132**(104324), 1–15 (2021)
17. Wang, D., et al.: Human-AI collaboration in data science: exploring data scientists' perceptions of Automated AI. In: Proceedings of the ACM on Human-Computer Interaction 3(CSCW), pp. 211:1–211:24 (2019)
18. Waring, J., Lindvall, C., Umeton, R.: Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif. Intell. Med. **104**(101822), 1–12 (2020)
19. World Health Organization: ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision, 2nd edn. World Health Organization (2004)

Appendix 6

[VI]

M. Bertl, M. Shahin, P. Ross, and D. Draheim. Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 826–833. Association for Computing Machinery, 2023

Finding Indicator Diseases of Psychiatric Disorders in BigData using Clustered Association Rule Mining

Markus Bertl
Department of Health Technologies,
Tallinn University of Technology
Tallinn, Estonia
mbertl@taltech.ee

Mahtab Shahin
Information System Group,
Tallinn University of Technology
Tallinn, Estonia
mahtab.shahin@taltech.ee

Peeter Ross
Department of Health Technologies,
Tallinn University of Technology
Tallinn, Estonia
East Tallinn Central Hospital
Tallinn, Estonia
peeter.ross@taltech.ee

Dirk Draheim
Information System Group,
Tallinn University of Technology
Tallinn, Estonia
dirk.draheim@taltech.ee

ABSTRACT

Psychiatric disorders represent critical non-communicable diseases of the 21st century and are ranked as the leading cause of years lived with disabilities. Nevertheless, data that could be used to improve our understanding of psychiatric diseases remain underutilized. In this research, we apply clustered association rule mining to find comorbidities and indicator diseases for patients with psychiatric illnesses. The model was trained with health insurance billing data from 60,115 patients with a total of 904,821 ICD-10 coded diseases. Nine association rules were found without clustering, 40 with clustering of F diagnoses. The approach proves suitable for further use in the implementation of indicator-based digital decision support systems in psychiatry.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → **Decision support systems**; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *Medical records*; *Medical technologies*;

KEYWORDS

Association rule mining (ARM), decision support systems (DSS), machine learning (ML), artificial intelligence (AI), explainable artificial intelligence (XAI), health insurance data, electronic health record (EHR), psychiatry, indicator diseases, comorbidity

ACM Reference Format:
Markus Bertl, Mahtab Shahin, Peeter Ross, and Dirk Draheim. 2023. Finding Indicator Diseases of Psychiatric Disorders in BigData using Clustered Association Rule Mining. In *Proceedings of ACM SAC Conference (SAC'23).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3555776.3577594

SAC'23, March 27 – March 31, 2023, Tallinn, Estonia

ACM ISBN 978-1-4503-9517-5/23/03…$15.00
https://doi.org/10.1145/3555776.3577594

1 INTRODUCTION

Psychiatric disorders represent critical non-communicable diseases of the 21st century and are ranked as the leading cause of years lived with disabilities [36]. Additionally, diagnosis accuracy is low [1, 23]. As one example, only 31% of patients with bipolar disorder are correctly diagnosed [33] – a diagnosis which takes 5.7 years on average [25]. The situation with depression looks similar [24]. Psychiatric diseases not only affect our health but also leave an impact from a cost perspective. In Europe alone, psychiatric disorders accounted for EUR 461 billion in healthcare costs [13].

With such a high impact on our wellbeing, as well as on society as a whole, a clear understanding of the factors that contribute to and/or indicate such diseases is crucial. Therefore, this research shows the potential of applying association rule mining to find the comorbidities of patients with psychiatric illnesses. Comorbidities are diseases that often co-occur with a primary condition. With hard-to-diagnose illnesses especially (such as psychiatric illnesses), a comorbidity could be used as an indicator that a currently undiagnosed root cause is present. If a patient has not yet been diagnosed with a psychiatric disease but suffers from diseases that normally co-occur with a psychiatric disease, a mental disorder could be the root cause.

Based on that principle, this paper shows an ARM-based approach for mining indicator diseases for psychiatry which could be used further for building digital decision support systems (DDSSs) to enable earlier diagnosis of mental health disorders. DDSSs based on health insurance claims data and machine learning techniques are nothing new [6]. However, despite their potential, they have not yet found their way to everyday clinical practice [5, 8]. The reasons for this are low user acceptance by health professionals [20] and the lack of explainability [4] resulting from the black-box character of many machine learning algorithms.

Association rule mining (ARM) can mitigate these problems. First introduced in 1993 [2], ARM remains one of the most popular methods of knowledge discovery [17]. Its history lies in finding patterns in transactional data, such as analysing which products in a shop sell best together (market basket analysis). In more general terms, an association rule like $A \Rightarrow B$ describes that in a dataset of transactions (D), a given transaction T containing itemset A is likely to also contain itemset B. The reliability of the association rule is expressed as confidence. Confidence is the percentage of transactions containing A and B compared to the total number of transactions containing A. In the market basket example from above, an association rule $(x_1, x_2) \Rightarrow (y_1)$ with a confidence of $C\%$ would mean that a customer who buys products x_1 and x_2 will also buy y_1 with a probability of $C\%$.

Due to its easily understandable and explainable nature, ARM could be a potentially good fit for DDSS. This research therefore investigates the following research question: What benefit does association rule mining bring to Digital Decision Support Systems in psychiatry?

The rest of this paper is organized as follows. Background and related work are detailed in Sect. 2, followed by a description of the used data and our methodology in Sect. 3. The experimental results from implementation are presented in Sect. 4. The validation approach of our results is described in Sect. 5. Sect. 6 discusses the results and their limitations and Sect. 7 concludes the paper.

2 BACKGROUND

Evidence indicates that those with psychiatric disorders compared to healthy subjects have a functional impairment that causes significant distress and poorer quality of life. [27]. Research studies provide massive volumes of heterogeneous data that are too complex and voluminous to be processed and analysed by traditional methods. We need inductive approaches like data mining methods to generate knowledge with a small number of cause-effect mechanisms.

Data mining is becoming increasingly popular for providing a deeper understanding of medical data, including disease pathogenesis and treatment, leading to discoveries from medical datasets that conventional methods are unable to process [18]. Data mining algorithms can be sorted into two main categories: supervised and unsupervised learning [15]. Supervised learning algorithms like classification or regression predict the response values for a particular outcome. Unsupervised learning algorithms describe data form and hidden structure using clustering and anomaly detection methods. One of the main unsupervised algorithms is association rule mining, which is widely used in various areas, one of which being medicine and health care [26, 29]. Association rule mining (ARM) finds associations and correlations throughout large sets of data and provides information in the form of 'if-then' probabilistic statements [32].

Several algorithms have been proposed for the generation of association rules [2]. Apriori is a well-known primary ARM algorithm for the extraction of frequent items in a set of transactions. First introduced by Agrawal et al.[2], Sharma and Om [30] used this algorithm for the early diagnosis and treatment of oral cancer. Karabatak and Ince [19] proposed an expert system for the detection of breast cancer based on association rules and neural networks. Chen et al. [10] used ARM to detect possible side effects due to exposure time to drugs during pregnancy, which resulted in the discovery of novel information. A few studies in psychiatry applied the association rule mining approach for data extraction. Shen et al. [31] used ARM on the Taiwan National Health Insurance Research Database to explore associations among comorbidities of borderline personality disorder. In another study, Schweikert et al. [28] applied Combinatorial Fusion Analysis (CFA) and ARM to explore the relationship between autism prevalence and lead mercury concentration, which led to a deeper understanding of its pathogenesis. Chen et al. [9] applied ARM to analyse the association among two or more comorbid diseases of attention deficit hyperactivity disorder. Hasanpour et al. [16] used the Apriori algorithm on an obsessive-compulsive disorder treatment response dataset of Iranian patients to explore the most significant factors contributing to the treatment response. Leejin and Sungmin [22] applied ARM on Korean National Health Insurance Data to extract comorbidities of attention-deficit hyperactivity disorder.

Previous psychiatry-related studies have focused on examining the strengths of associations [35] and classifying patients from real-world clinic databases [22]. However, these methods focus on examining symptom patterns. The current paper focuses on using pattern mining techniques known as clustered ARM to provide a descriptive approach for extracting symptom rules. Most research only tackles a specific disease. Due to our large dataset, which covers many psychiatric diseases of almost the entire population of Estonia, we intend to show that this approach can be generalized for psychiatry as a whole. No previous studies have focused on the analysis of comorbidities and indicator diseases for patients with psychiatric illnesses using ARM. This study aims to discover the hidden relationships between patients with psychiatric illnesses, which can aid medical decision-making for a faster, more accurate diagnosis. Moreover, to the best of the authors' knowledge, the collected dataset is one of the most extensive datasets covering psychiatric disorders. This makes our research results valuable not only for medical professionals, but also for the computer science community by showing how clustered ARM can be applied to BigData.

3 METHOD

This research is based on association rule mining, a well-recognized approach in data science and knowledge discovery [34][21]. We extend this well-established approach by applying ARM to hierarchically clustered data, thereby demonstrating not only the feasibility of this approach for this BigData application scenario (especially concerning the variety and veracity of the data), but also its potential for digital decision support.

3.1 Data

The data used were collected from the Estonian Health Insurance Fund (EHIF). The EHIF manages healthcare expenses reimbursement. Their digital system was introduced in 2001 as an addition to the paper-based process. Since 2005, all reimbursement claims and prescriptions needed to be submitted electronically.

As of today, the process of data collection from the EHIF is part of there reimbursement process of healthcare providers. Medical professionals fill in the case history and demographic data in a structured electronic medical record system. Then they compose a discharge letter or outpatient summary where the activities are coded based on ICD-10 codes. The International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) is an international standardized coding system for reporting diseases [37]. Each ICD-10 code is structured based on an alpha character called a chapter and two digits describing the category of the disease, followed by a dot and further digits representing more details such as cause, location, severity or other clinical information (sub-categories). As one example, F32.2 codes for major depressive disorder, single episode, severe without psychotic features. F stands for mental and behavioral disorders, F30-F39 code mood [affective] disorders, where F32 is the category major depressive disorder, single episode. The '.2' at the end specifies the severity. This information is then sent to the EHIF where it is automatically quality checked, and a random sample set of cases is manually validated before the reimbursement process is started. After the checks, the data is saved in the EHIF data warehouse (DWH). The data is then used by the medical statisticians of the EHIF, researchers, political or healthcare policy decision-makers, or other governmental authorities. Based on the role of the data requestor, the data is available either in personalized form with patient identifiers, or anonymized.

Our dataset was queried from the EHIF DWH and includes information on gender, birth year, ICD-10 coded diagnosis, and diagnoses year and month from 60,115 adults (18 years or above) with a total of 904,821 diagnoses in 2018 and 2019. The data consists of all publicly insured people in Estonia with a depression diagnosis, either single episode (F32), or recurrent (F33), and an equally-sized random sample of people with other psychiatric disorders. The percentage of insured people in Estonia is above 93.63% [12], so we are confident that our dataset is representative for the whole population. Since we only obtained and analyzed anonymized data, no ethics approval was required.

To decrease the dimensionality of the data, the ICD codes were clustered by category level by removing the digits preceding the dot. Additionally, chapter XXI "Factors influencing health status and contact with health services" (categories Z00-Z99) was removed. Codes in this category are provided for occasions when circumstances other than a disease, injury or external cause is recorded [37]. Examples are examinations, prescriptions or diagnostic tests.

For better readability in this publication, the ICD-10 codes were mapped to their text representation using the official ICD-10 XML mapping files provided by the Estonian Health Insurance Fund [14].

3.2 Clustered Association Rule Mining

In this research, we are using the Apriori algorithm to mine the rules. First introduced in 1994, it has become the standard algorithm for association rule mining [3]. The algorithm was first designed to identify frequent itemsets in transnational data. Given a threshold C, it finds frequent itemsets that are subsets of at least C transactions using a bottom-up approach based on breadth-first search and hash trees [3]. An item set X of length k is frequent if and only if every subset of X, having a length of $k-1$, is also frequent. Rules are then ranked according to interest metrics. The most basic measurement of interestingness is support [2], also called frequency constraints. It describes the fraction of all the transactions in which an item set occurs divided by the total number of transactions. Other measurements of interest are derived from support. For each rule mined in this research, we calculated:

- $support(A \Rightarrow C) = support(A + C)$
  range: [0, 1]
- $confidence(A \Rightarrow C) = support(A + C)/support(A)$
  range [0, 1]
- $lift(A \Rightarrow C) = confidence(A \Rightarrow C)/support(C)$
  range: [0, inf]
- $leverage(A \Rightarrow C) = support(A \Rightarrow C) - support(A) * support(C)$
  range: [-1, 1]
- $conviction = [1 - support(C)]/[1 - confidence(A \Rightarrow C)]$
  range: [0, inf]

Traditional ARM algorithms often cannot produce good results with high-volume and high-dimensional data because ARM is not able to generalize. In our example, significant associations with diseases that would be visible when looking at aggregated groups (e.g., mood disorders in general, not only depression) might not be visible for individual psychiatric disorders. Clustered association rule mining can be used to solve this problem. With clustered ARM, the data are clustered (either by ML algorithms like K-means or – as in our example – based on the existing hierarchic data structure of ICD-10) before ARM is applied.

3.3 Development Environment, Libraries & Configuration

Our development environment consisted of *JupyterLab 3.0.12* with *Python 3*, *Pandas 1.2.3* and *PyCaret 2.3.5* deployed on a virtualized Ubuntu server with an 8 Core Intel Xeon E5-2650 v2 CPU @ 2.6 GHz and 64 GB RAM.

4 RESULTS

We used two data pre-processing approaches for the association rule mining. For the first approach, the original data were taken. Because of the high dimensionality of the data (1355 variables/different diseases), we decided to assign all psychiatric diagnoses to one cluster (F) before the rule mining to reduce the dimensionality and get better results. Since we were only interested in finding possible indicator diseases for psychiatric disorders, all association rules with no psychiatric disease as a consequence were removed for better visibility. No further post-processing was applied. In order to derive a relevant set of rules, we iteratively applied different interestingness constraints. The final constraints used are mentioned in each section.

4.1 Association Rules without Clustering

In total, 80 association rules were mined, of which nine had a psychiatric disease as a consequence. Table 1 shows the association rules mined from the original dataset without any clustering applied. Table 2 translates the ICD codes to the corresponding description.

To reduce the set of association rules, interestingness constraints of 0.0155 as support threshold and 1.25 as lift threshold were used.

4.2 Association Rules with Clustering

In total, 80 association rules were minded, of which 40 had a psychiatric disease as consequence. Table 3 shows the association rules mined with previous clustering of all psychiatric diseases into one category (F). Table 4 translates the ICD codes to the corresponding description. To reduce the set of association rules, interestingness constraints of 0.025 as support threshold and 1 as lift threshold were used.

5 VALIDATION

To ensure the clinical plausibility and validity of the presented results, we verified the found rules with both medical experts and medical literature. We interviewed two domain experts from psychiatry and psychology and presented our research approach and the found rules. Both experts positively attested the results obtained. Additionally, the presented results have some overlaps with medical research on comorbidities of psychiatric diseases [11], which increases the external validity of our findings.

6 DISCUSSION

Since comorbidities are often derived from clinical guidelines, which are mostly based on studies from bigger western countries, comorbidity research is often biased towards bigger countries. This is especially relevant to psychiatry because it has a strong cultural component; indicator diseases might vary between countries. Our approach could be used to mine more targeted, country-specific comorbidities than currently available in literature. Additionally, there is currently no standardized process on how comorbidities are mined in health care. Most were found by manual statistical analysis or clinical guidelines. For new diseases like COVID-19 especially, our approach can be used to vastly mine comorbidities and refine them as soon as variables like health policies or the disease itself change. This is especially relevant since clinical guidelines usually take a long time to adapt.

In future research, our association rules could be used as input for a DDSS that can alert when a patient is frequently diagnosed with diseases that correlate to psychiatric diseases. Using this approach, possible underlying psychiatric diseases could be diagnosed faster. Currently, deep learning algorithms are often used for the cognition of such DDSSs. Deep neural networks not only deal with high-dimensional data in health care, they also learn from complex time-series data. This is especially helpful for diseases with a high inter-patient variety. However, techniques like deep learning are by default not explainable. This black-box character has been identified as a barrier to the adoption of DDSSs [7]. With an indicator disease-based DDSS, the diagnoses leading to a DDSS alert can be made transparent to medical professionals. The explainability of ARM could therefore help raise the currently low user acceptance of DDSSs in psychiatry. Additionally, DDSSs based on deep neural networks are highly computationally intensive. The computational costs of mining ARM are much cheaper. While deep learning often requires high-performance computing, or at least GPU-powered training devices, the ARM approach demonstrated ran on the above-mentioned standard server in under a minute.

For the reasons set out above, we see high potential for our ARM approach not only in further advancing public health research by allowing timely, country-specific comorbidity mining, but also as a rule basis for digital decision support.

In terms of limitations, it is important to note that health insurance billing data holds a certain amount of bias. One example is that they cover only the insured population, so it might not be possible to completely generalize the results for everyone. For the dataset we used, this can be disregarded, since the percentage of insured people in Estonia is around 93.63% [12]. However, the data are collected for claims management. Therefore, it might not represent the medical truth completely. Especially since private consultations, which frequently happen in psychiatry, are typically not recorded by those systems. Furthermore, the algorithm has certain limitations. Since association rule mining often produces quite a large set of rules, of which typically only a small subset is relevant, it relies on picking the correct interesting rules. This is mostly done manually, with the support of different measurements of interestingness. Manual work is prone to confirmation bias. Additionally, ARM works based on frequencies. It does not – and cannot – take into account medical cause and effect. Especially with healthcare data, where some diseases are widespread and others rare, measurements of interestingness can deviate from normal ranges. Additionally, the large size of the dataset we used in this research contributes to that effect. An additional point of limitation, and an area into which we encourage further research, is clinical validation. In order to design and develop a working decision support system for psychiatry, the found rules need to be integrated not only into a rule engine but also into the clinical process. This requires further research on where a DDSS like this would bring the most potential as well as further research on the benefits and validity thereof.

7 CONCLUSION

This work presented a knowledge mining example based on association rules for extracting indicator diseases of psychiatric disorders. Compared to other approaches that utilize healthcare data for patient benefits, such as deep learning-based decision support, ARM excels by offering complete explainability and low computational costs. In terms of medical validity, the found association rules match clinical guidelines in psychiatry, which indicates the reliability of ARM. This research has therefore demonstrated the usefulness of association rule mining for comorbidity extraction and indicator disease mining based on health insurance billing data. A possible use case for this is public health research in order to allow timely, country-specific comorbidity mining. The second area of application could be the quick and easy creation of a knowledge-base for a DDSS, which can alert when a psychological condition might be present if high correlating diseases are frequently diagnosed.

ACKNOWLEDGMENTS

This work in the project "ICT programme" was partially supported by the European Union through European Social Fund.

Table 1: Association rules of ICD-10 codes without F-clustering

#	antecedents	consequents	ant. sup.	con. sup.	support	confidence	lift	leverage	conviction
1	(I11)	(F51)	0.1602	0.0586	0.0168	0.1051	1.7935	0.0074	1.0519
2	(M17)	(F33)	0.0615	0.1861	0.0157	0.2553	1.3718	0.0043	1.0929
3	(K21)	(F33)	0.0639	0.1861	0.0159	0.2490	1.3379	0.0040	1.0837
4	(G47)	(F33)	0.0778	0.1861	0.0193	0.2479	1.3321	0.0048	1.0822
5	(R10)	(F41)	0.0822	0.1976	0.0215	0.2621	1.3264	0.0053	1.0873
6	(K21)	(F41)	0.0639	0.1976	0.0167	0.2604	1.3178	0.0040	1.0849
7	(G47)	(F41)	0.0778	0.1976	0.0197	0.2537	1.2839	0.0044	1.0751
8	(N30)	(F41)	0.0746	0.1976	0.0187	0.2506	1.2682	0.0039	1.0706
9	(G47)	(F32)	0.0778	0.2110	0.0206	0.2644	1.2531	0.0042	1.0725

Table 2: Association rules without F-clustering (mapping table)

#	antecedents	count	consequents	count
1	Hypertensive heart disease	33624	Nonorganic sleep disorders	10956
2	Gonarthrosis [arthrosis of the knee]	9471	Recurrent depressive disorder	59941
3	Gastro-oesophageal reflux disease	7350	Recurrent depressive disorder	59941
4	Sleep disorders	14677	Recurrent depressive disorder	59941
5	Abdominal and pelvic pain	6755	Other anxiety disorders	42990
6	Gastro-oesophageal reflux disease	7350	Other anxiety disorders	42990
7	Sleep disorders	14677	Other anxiety disorders	42990
8	Cystitis	6711	Other anxiety disorders	42990
9	Sleep disorders	14677	Depressive episode	53034

Table 3: Association rules of ICD-10 codes with F clustering

#	antecedents	consequents	ant. sup.	con. sup.	support	confidence	lift	leverage	conviction
1	(I11)	(F, M54)	0.1602	0.1584	0.0324	0.2023	1.2771	0.0070	1.0550
2	(I10)	(F, M54)	0.1362	0.1584	0.0275	0.2020	1.2753	0.0059	1.0546
3	(M54)	(F, I10)	0.2066	0.1059	0.0275	0.1332	1.2578	0.0056	1.0315
4	(M54)	(I11, F)	0.2066	0.1248	0.0324	0.1568	1.2564	0.0066	1.0380
5	(M54)	(F, J06)	0.2066	0.1359	0.0352	0.1704	1.2539	0.0071	1.0416
6	(J06)	(F, M54)	0.1834	0.1584	0.0352	0.1920	1.2121	0.0062	1.0416
7	(G47)	(F)	0.0778	0.7474	0.0625	0.8036	1.0752	0.0044	1.2862
8	(I11, M54)	(F)	0.0406	0.7474	0.0324	0.7980	1.0677	0.0021	1.2505
9	(K21)	(F)	0.0639	0.7474	0.0508	0.7947	1.0633	0.0030	1.2304
10	(I10, M54)	(F)	0.0348	0.7474	0.0275	0.7918	1.0594	0.0015	1.2133
11	(I49)	(F)	0.0340	0.7474	0.0268	0.7892	1.0560	0.0014	1.2133
12	(H25)	(F)	0.0450	0.7474	0.0353	0.7843	1.0494	0.0017	1.1713
13	(E78)	(F)	0.0425	0.7474	0.0333	0.7842	1.0492	0.0016	1.1704
14	(J45)	(F)	0.0496	0.7474	0.0389	0.7842	1.0492	0.0018	1.1703
15	(M17)	(F)	0.0615	0.7474	0.0481	0.7820	1.0464	0.0021	1.1589
16	(H52)	(F)	0.0862	0.7474	0.0674	0.7818	1.0460	0.0030	1.1576
17	(M51)	(F)	0.0516	0.7474	0.0403	0.7813	1.0453	0.0017	1.1550
18	(K29)	(F)	0.0746	0.7474	0.0583	0.7812	1.0453	0.0025	1.1546
19	(R51)	(F)	0.0343	0.7474	0.0268	0.7809	1.0448	0.0011	1.1529
20	(G44)	(F)	0.0357	0.7474	0.0279	0.7794	1.0429	0.0011	1.1452
21	(I11)	(F)	0.1602	0.7474	0.1248	0.7790	1.0423	0.0051	1.1432
22	(M15)	(F)	0.0355	0.7474	0.0276	0.7780	1.0409	0.0011	1.1378
23	(N39)	(F)	0.0471	0.7474	0.0366	0.7779	1.0408	0.0014	1.1373
24	(I10)	(F)	0.1362	0.7474	0.1059	0.7770	1.0397	0.0040	1.1330
25	(R10)	(F)	0.0822	0.7474	0.0637	0.7753	1.0374	0.0023	1.1244
26	(E11)	(F)	0.0489	0.7474	0.0378	0.7731	1.0343	0.0013	1.1130
27	(R07)	(F)	0.0334	0.7474	0.0257	0.7695	1.0296	0.0007	1.0961
28	(H61)	(F)	0.0352	0.7474	0.0270	0.7687	1.0285	0.0007	1.0920
29	(M79)	(F)	0.0734	0.7474	0.0563	0.7677	1.0271	0.0015	1.0873
30	(M54)	(F)	0.2066	0.7474	0.1584	0.7665	1.0256	0.0040	1.0820
31	(J20)	(F)	0.0693	0.7474	0.0531	0.7661	1.0250	0.0013	1.0799
32	(J06, M54)	(F)	0.0462	0.7474	0.0352	0.7626	1.0203	0.0007	1.0641
33	(E03)	(F)	0.0443	0.7474	0.0336	0.7580	1.0141	0.0005	1.0437
34	(M25)	(F)	0.0766	0.7474	0.0579	0.7559	1.0113	0.0006	1.0347
35	(M75)	(F)	0.0407	0.7474	0.0306	0.7521	1.0063	0.0002	1.0191
36	(H10)	(F)	0.0590	0.7474	0.0444	0.7517	1.0057	0.0003	1.0173
37	(H40)	(F)	0.0389	0.7474	0.0293	0.7514	1.0053	0.0002	1.0160
38	(N30)	(F)	0.0746	0.7474	0.0559	0.7497	1.0030	0.0002	1.0090
39	(N95)	(F)	0.0546	0.7474	0.0409	0.7490	1.0022	0.0001	1.0064
40	(J04)	(F)	0.0391	0.7474	0.0293	0.7483	1.0012	0.0000	1.0036

Table 4: Association rules with F-clustering (mapping table)

#	antecedents	count	consequents	count
1	(Hypertensive heart disease)	(33624)	(F, Dorsalgia)	(254113, 24473)
2	(Essential (primary) hypertension)	(23223)	(F, Dorsalgia)	(254113, 24473)
3	(Dorsalgia)	(24473)	(F, Essential (primary) hypertension)	(254113, 23223)
4	(Dorsalgia)	(24473)	(Hypertensive heart disease, F)	(33624 ,254113)

Continued on next page

Table 4 – continued from previous page

#	antecedents	count	consequents	count
5	(Dorsalgia)	(24473)	(F, Acute upper respiratory infections of multiple and unspecified sites)	(254113, 15507)
6	(Acute upper respiratory infections of multiple and unspecified sites)	(15507)	(F, Dorsalgia)	(254113, 24473)
7	(Sleep disorders)	(14677)	(F)	(254113)
8	(Hypertensive heart disease, Dorsalgia)	(33624, 24473)	(F)	(254113)
9	(Gastro-oesophageal reflux disease)	(7350)	(F)	(254113)
10	(Essential (primary) hypertension, Dorsalgia)	(23223, 24473)	(F)	(254113)
11	(Other cardiac arrhythmias)	(3849)	(F)	(254113)
12	(Senile cataract)	(4860)	(F)	(254113)
13	(Disorders of lipoprotein metabolism and other lipidaemias)	(4226)	(F)	(254113)
14	(Asthma)	(7922)	(F)	(254113)
15	(Gonarthrosis [arthrosis of knee])	(9471)	(F)	(254113)
16	(Disorders of refraction and accommodation)	(6118)	(F)	(254113)
17	(Other intervertebral disc disorders)	(7234)	(F)	(254113)
18	(Gastritis and duodenitis)	(8235)	(F)	(254113)
19	(Headache)	(2610)	(F)	(254113)
20	(Other headache syndromes)	(3839)	(F)	(254113)
21	(Hypertensive heart disease)	(33624)	(F)	(254113)
22	(Polyarthrosis)	(4423)	(F)	(254113)
23	(Other disorders of urinary system)	(4173)	(F)	(254113)
24	(Essential (primary) hypertension)	(23223)	(F)	(254113)
25	(Abdominal and pelvic pain)	(6755)	(F)	(254113)
26	(Type 2 diabetes mellitus)	(11880)	(F)	(254113)
27	(Pain in throat and chest)	(2384)	(F)	(254113)
28	(Other disorders of external ear)	(2513)	(F)	(254113)
29	(Other soft tissue disorders, not elsewhere classified)	(6502)	(F)	(254113)
30	(Dorsalgia)	(24473)	(F)	(254113)
31	(Acute bronchitis)	(5068)	(F)	(254113)
32	(Acute upper respiratory infections of multiple and unspecified sites, Dorsalgia)	(15507, 24473)	(F)	(254113)
33	(Other hypothyroidism)	(6933)	(F)	(254113)
34	(Other joint disorders, not elsewhere classified)	(6618)	(F)	(254113)
35	(Shoulder lesions)	(5065)	(F)	(254113)
36	(Conjunctivitis)	(4726)	(F)	(254113)
37	(Glaucoma)	(8829)	(F)	(254113)
38	(Cystitis)	(6711)	(F)	(254113)
39	(Menopausal and other perimenopausal disorders)	(4955)	(F)	(254113)
40	(Acute laryngitis and tracheitis)	(2870)	(F)	(254113)

REFERENCES

[1] Ahmed Aboraya, Eric Rankin, Cheryl France, Ahmed El-Missiry, and Collin John. 2006. The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont)* 3, 1 (2006), 41.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 207–216.

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Citeseer, 487–499.

[4] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 559–560.

[5] Markus Bertl, Janek Metsallik, and Peeter Ross. 2022. A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Frontiers in Psychiatry* 13 (2022). https://doi.org/10.3389/fpsyt.2022.923613

[6] Markus Bertl, Peeter Ross, and Dirk Draheim. 2021. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *International Conference on Database and Expert Systems Applications*. Springer, 104–111.

[7] Markus Bertl, Peeter Ross, and Dirk Draheim. 2021. Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors Systematic AI Support in the Healthcare Sector. (2021). https://doi.org/10.13140/RG.2.2.17159.52646/1

[8] Markus Bertl, Peeter Ross, and Dirk Draheim. 2022. A survey on AI and Decision Support Systems in Psychiatry - Uncovering a Dilemma. *Expert Systems with Applications* 202 (2022), 117464. https://doi.org/10.1016/j.eswa.2022.117464

[9] Tzeng-Ji Chen, Li-Fang Chou, and Shinn-Jang Hwang. 2003. Application of a data-mining technique to analyze coprescription patterns for antacids in Taiwan. *Clinical therapeutics* 25, 9 (2003), 2453–2463.

[10] Yu Chen, Lars Henning Pedersen, Wesley W Chu, and Jorn Olsen. 2007. Drug exposure side effects from mining pregnancy data. *ACM SIGKDD Explorations Newsletter* 9, 1 (2007), 22–29.

[11] William W. Eaton. 2007. *Medical and psychiatric comorbidity over the course of life*. American Psychiatric Pub.

[12] Estonian National Institute for Health Development. 2020. RA02: Residents with health insurance and health insurance coverage by sex and county-Tervisestatistika ja terviseuuringute andmebaas. https://statistika.tai.ee/pxweb/en/Andmebaas/Andmebaas__04THressursid__12Ravikindlustatud/RA02.px/, Last accessed on 2022-03-04.

[13] Anders Gustavsson, Mikael Svensson, Frank Jacobi, Christer Allgulander, Jordi Alonso, Ettore Beghi, Richard Dodel, Mattias Ekman, Carlo Faravelli, Laura Fratiglioni, et al. 2011. Cost of disorders of the brain in Europe 2010. *European neuropsychopharmacology* 21, 10 (2011), 718–779.

[14] Haigekassa. 2021. Website of the Estonian Health Insurance Fund. http://haigekassa.ee

[15] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[16] Hesam Hasanpour, Ramak Ghavamizadeh Meibodi, Keivan Navi, Jamal Shams, Sareh Asadi, and Abolhassan Ahmadiani. 2019. Fluvoxamine treatment response prediction in obsessive-compulsive disorder: association rule mining approach. *Neuropsychiatric disease and treatment* 15 (2019), 895.

[17] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. 2000. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter* 2, 1 (2000), 58–64.

[18] Zan Huang, Jiexun Li, Hua Su, George S Watts, and Hsinchun Chen. 2007. Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining. *Decision Support Systems* 43, 4 (2007), 1207–1225.

[19] Murat Karabatak and M Cevdet Ince. 2009. An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications* 36, 2 (2009), 3465–3469.

[20] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

[21] Trupti A Kumbhare and Santosh V Chobe. 2014. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies* 5, 1 (2014), 927–930.

[22] KIM Leejin and Sungmin Myoung. 2018. Comorbidity study of attention-deficit hyperactivity disorder (ADHD) in children: applying association rule mining (ARM) to Korean National Health Insurance Data. *Iranian journal of public health* 47, 4 (2018), 481.

[23] Elizabeth A McGlynn, Steven M Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A Kerr. 2003. The quality of health care delivered to adults in the United States. *New England journal of medicine* 348, 26 (2003), 2635–2645.

[24] Alex J Mitchell, Amol Vaze, and Sanjay Rao. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374, 9690 (2009), 609–619.

[25] Paolo Lucio Morselli and Rodney Elgie. 2003. GAMIAN-Europe*/BEAM survey I–global analysis of a patient questionnaire circulated to 3450 members of 12 European advocacy groups operating in the field of mood disorders. *Bipolar Disorders* 5, 4 (2003), 265–278.

[26] Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen. 2013. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications* 40, 4 (2013), 1086–1093.

[27] Samuli I Saarni, Jaana Suvisaari, Harri Sintonen, Sami Pirkola, Seppo Koskinen, Arpo Aromaa, and Jouko Lönnqvist. 2007. Impact of psychiatric disorders on health-related quality of life: general population survey. *The British journal of psychiatry* 190, 4 (2007), 326–332.

[28] Christina Schweikert, Yanjun Li, David Dayya, David Yens, Martin Torrents, and D Frank Hsu. 2009. Analysis of autism prevalence and neurotoxins using combinatorial fusion and association rule mining. In *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*. IEEE, 400–404.

[29] Mahtab Shahin, Wissem Inoubli, Syed Attique Shah, Sadok Ben Yahia, and Dirk Draheim. 2021. Distributed scalable association rule mining over COVID-19 data. In *International Conference on Future Data and Security Engineering*. Springer, 39–52.

[30] Neha Sharma and Hari Om. 2014. Extracting significant patterns for oral cancer detection using apriori algorithm. *Intelligent Information Management* 2014 (2014).

[31] Cheng-Che Shen, Li-Yu Hu, and Ya-Han Hu. 2017. Comorbidity study of borderline personality disorder: applying association rule mining to the Taiwan national health insurance research database. *BMC medical informatics and decision making* 17, 1 (2017), 1–10.

[32] A Mi Shin, In Hee Lee, Gyeong Ho Lee, Hee Joon Park, Hyung Seop Park, Kyung Il Yoon, Jung Jeung Lee, and Yoon Nyun Kim. 2010. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthcare informatics research* 16, 2 (2010), 77–81.

[33] Tanvir Singh and Muhammad Rajput. 2006. Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)* 3, 10 (2006), 57.

[34] Surbhi K Solanki and Jalpa T Patel. 2015. A survey on association rule mining. In *2015 fifth international conference on advanced computing & communication technologies*. IEEE, 212–216.

[35] Yueh-Ming Tai and Hung-Wen Chiu. 2009. Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. *International journal of medical informatics* 78, 12 (2009), e75–e83.

[36] Hans-Ulrich Wittchen, Frank Jacobi, Jürgen Rehm, Anders Gustavsson, Mikael Svensson, Bengt Jönsson, Jes Olesen, Christer Allgulander, Jordi Alonso, Carlo Faravelli, et al. 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European neuropsychopharmacology* 21, 9 (2011), 655–679.

[37] World Health Organization. 2004. *ICD-10 : International statistical classification of diseases and related health problems : tenth revision* (2nd ed ed.). World Health Organization.

Appendix 7

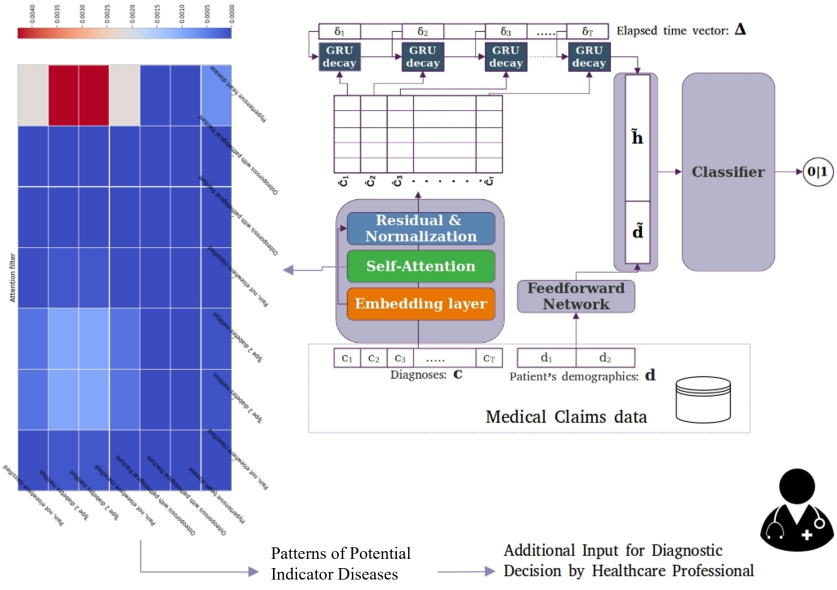
[VII]

M. Bertl, N. Bignoumba, P. Ross, S. B. Yahia, and D. Draheim. Evaluation of Deep Learning-based Depression Detection using Medical Claims Data. *SSRN*, 2023

Graphical Abstract

Evaluation of Deep Learning-based Depression Detection using Medical Claims Data

Markus Bertl, Nzamba Bignoumba, Peeter Ross, Sadok Ben Yahia, Dirk Draheim



Highlights

Evaluation of Deep Learning-based Depression Detection using Medical Claims Data

Markus Bertl, Nzamba Bignoumba, Peeter Ross, Sadok Ben Yahia, Dirk Draheim

- We evaluate common AI models for depression detection based on a real-world dataset containing EHR data of 812,853 patients.
- We propose our own Att-GRU-decay model, which outperforms the current state of the art.
- Our Att-GRU-decay model detects depression with 0.974 AUPRC, 0.999 specificity and 0.944 sensitivity.
- The included self-attention layer uncovers disease patterns indicating depression.
- We finally contribute by showing how the Att-GRU-decay model can be used in a GP setting to improve the diagnosis of depression.

Evaluation of Deep Learning-based Depression Detection using Medical Claims Data

Markus Bertl[a], Nzamba Bignoumba[b], Peeter Ross[a], Sadok Ben Yahia[b,c], Dirk Draheim[b]

[a] *Department of Health Technologies, Tallinn University of Technology, Akadeemia Tee 15A, Tallinn, 12618, Estonia*
[b] *Department of Software Science, Tallinn University of Technology, Akadeemia Tee 15A, Tallinn, 12618, Estonia*
[c] *University of Southern Denmark, Alsion 2, Sønderborg, 6400, Denmark*

Abstract

Human accuracy in diagnosing psychiatric disorders is still low. Even though digitizing health care leads to more and more data, the successful adoption of AI-based digital decision support (DDSS) is rare. One reason is that AI algorithms are often not evaluated based on large, real-world data. This research shows the potential of using deep learning on the medical claims data of 812,853 people between 2018 and 2022, with 26,973,943 ICD-10-coded diseases, to predict depression (F32 and F33 ICD-10 codes). The dataset used represents almost the entire adult population of Estonia. Based on these data, to show the critical importance of the underlying temporal properties of the data for the detection of depression, we evaluate the performance of non-sequential models (LR, FNN), sequential models (LSTM, CNN-LSTM) and the sequential model with a decay factor (GRU-Δt, GRU-decay). Furthermore, since explainability is necessary for the medical domain, we combine a self-attention model with the GRU decay and evaluate its performance. We named this combination Att-GRU-decay. After extensive empirical experimentation, our model (Att-GRU-decay), with an AUC score of 0.990, an AUPRC score of 0.974, a specificity of 0.999 and a sensitivity of 0.944, proved to be the most accurate. The results of our novel Att-GRU-decay model outperform the current state of the art, demonstrating the potential usefulness of deep learning algorithms for DDSS development. We further expand this by describing a possible application scenario of the proposed algorithm for depression screening in a general practitioner (GP) setting—not only to decrease healthcare costs, but also to improve the quality of care and

ultimately decrease people's suffering.

Keywords: Artificial intelligence (AI), decision support system (DSS), deep learning, machine learning (ML), depression, insurance data, medical claims data, psychiatry

1. Introduction

Psychiatric disorders, especially mood disorders such as depression, represent the critical non-communicable diseases of the 21^{st} century and are ranked as the leading cause of years lived with disabilities [1]. Unfortunately, these diseases are often diagnosed late or incorrectly [2, 3]. According to [4], depression is diagnosed by general practitioners with a sensitivity of 50.1% (95% CI: 41.3 to 59.0) and a specificity of 81.3% (95% CI: 74.5 to 87.3). Other research suggests that in the US, two-thirds of depression patients go undiagnosed [5]. Psychiatric diseases affect people's health and leave an impact from a cost perspective on a more global level. In Europe alone, psychiatric disorders accounted for EUR 461 billion in healthcare costs [6]. Of all psychiatric diseases, the economic costs of depression are among the highest. Other research suggests that the quality-adjusted life years (QUALYs) lost amount to $9,950 per citizen with undiagnosed depression [7].

The recommended method for diagnosing depression in 2023 is based on questionnaires and assessment scales from the previous century [8]. In psychology and psychiatry, medical professionals still rely on methods dating back to the 1960s [9, 10]. New technologies such as Artificial Intelligence (AI), especially deep learning, could potentially improve this situation by supporting medical professionals. The computer-based systems that use data to assist decision-making are called *digital decision support systems* (DDSS). AI-based DDSS for psychiatry is an active research field [11, 12]. However, research often does not make its way into clinical practice. One reason is that the data used to develop such systems are often unavailable or of bad quality [13]. More and more countries are applying a single public payer approach to health care, like Canada [14], Australia [15], the UK [16] or Estonia [15], meaning that a vast amount of medical claims data will be available in a central place. However, data are mostly used for claims management and rarely reused. This paper investigates which algorithm is best suited for building a deep learning-based DDSS for depression detection based on medical claims data. As one of the leaders of e-government [17], Estonia is a good starting

point for such research because lots of digital data are already available.

The foundation of the Estonian e-state is its digital identity system [18]. Each of the 1.3 million residents of Estonia has their own unique ID code. This allows for the creation of digital government services like online income tax declaration (used by 96% of people [19]), internet voting (used by 46.7% [20]) or e-prescription (used by 99% [21]) – e-health services especially profit from the Estonian e-state. One example is the collection of medical data. In Estonia, medical data are saved in two central places. The first is the e-health system called the Estonian Nationwide Health Information System (NHIS). The NHIS has been operational since 2008, allowing secure and trusted online access to medical data, prescriptions and medical images for virtually all Estonian residents. Instead of one big centralized database, the NHIS comprises several federated and mutually independent systems. One of them is the nationwide electronic health record (EHR) system. In the central EHR, patient data are saved based on international standards such as HL7 CDA[1], DICOM[2], LOINC[3], ICD-10[4] and SNOMED-CT[5].

The second place is the Estonian Health Insurance Fund (EHIF), which manages healthcare expense reimbursement. Their digital system was introduced in 2001 as an addition to the paper-based process. Since 2005, all reimbursement claims and prescriptions must be submitted electronically. As of today, the data collection process of the EHIF is part of their reimbursement process for healthcare providers. Medical professionals fill in the case history and demographic data in a structured electronic medical record system. Then they compose a discharge letter or an outpatient summary, where the activities are coded using ICD-10 codes. This information is then sent to the EHIF, where it is automatically quality-checked and a random sample set of cases is manually validated before the reimbursement process is initiated. After the checks, the data are saved in the EHIF data warehouse. The medical statisticians of the EHIF then use the data for research, political or healthcare policy decision-making or to supply information to other governmental authorities. Based on the role of the data requester, the data are available in either personalized form with patient identifiers or in

[1] `http://www.hl7.org/`
[2] `https://www.dicomstandard.org/`
[3] `https://loinc.org/`
[4] `https://icd.who.int/browse10/`
[5] `https://www.snomed.org/`

anonymized form. In the case of Estonia, a duplication of the data is also saved to the NHIS.

Since ICD codes are a concatenation of digits and alphabetic characters that convey information, they are considered categorical features. Thus, for analysis and prediction purposes involving ICD codes, we can benefit from the state-of-the-art machine and deep learning models dedicated to NLP tasks, such as [22, 23]. For our work, we leverage on a *self-attention layer*[6], which is a *transformer's sub-layer* [23], for efficient encoding of hidden relationships between diagnoses.

Since a single data modality (in our case, diagnosis) is usually not consistent enough for effective decision-making, it is common in the medical field to merge heterogeneous features or homogeneous features with different modalities, such as clinical text, demographics, images, or IoT sensor data [24, 25, 26, 27]. In the case of depression detection, due to its heterogeneity, i.e. different types of depression, several heterogeneous data are usually involved. Indeed, the high number of similarities that exist between some depression types make their classification complicated. Although considering several heterogeneous data may improve model accuracy, some are difficult to collect or biased with inconsistent patient responses. Assuming that diagnoses performed by a doctor are more trusted and accessible, we decided to combine them with demographic data for more accurate depression detection.

Medical events are recorded with their corresponding date in the patient's electronic health record (EHR). The recording date plays a crucial role in the clinical process: it allows practitioners to track the trajectory of the patient's health status over time to make appropriate decisions. The omission of this information for decision support will undoubtedly result in lower performance and make the model less realistic. It is therefore necessary to consider the sequence of medical events. Like many models that have been built for health care (or used a medical problem as a pilot case) [28, 29], we also consider the time intervals between consecutive diagnoses as an additional input for our model so that the process of detecting depression can rely more on recent diagnoses. As in [28, 30], we model this temporal aspect by incorporating a decay factor in the gated recurrent unit (GRU) [31]. Considering the time intervals between consecutive diagnoses as additional inputs and effectively incorporating them into the GRU's core via a decay factor sets our proposal

[6]The word layer can be used interchangeably with block, model or component.

apart from previous works on depression detection [32].

Depression is often addressed like a binary classification or multimodal logistic regression problem [33, 34]. Binary classification determines whether a patient suffers from depression, while multimodal logistic regression associates each type of depression with a probability score. The highest scoring type is then selected as the diagnosis. Although multimodal logistic regression has the advantage of learning the distribution of each depression type mutually, it faces the problem of imbalanced class distribution. Moreover, only patients who suffer from depression are studied. As we want to minimize imbalanced class problems, detecting patients suffering from depression and those not suffering, we choose the binary classification approach with class-weighting factors. Unlike our predecessors, here are additional aspects that we considered:

- Using a *self-attention layer* to effectively learn hidden relationships between diagnoses to better represent patient health status.
- Weighting the significance of diagnoses based on their corresponding record date so that the model can rely more on recently made diagnoses.
- Using *weighted binary cross-entropy* as the loss function to deal with the imbalanced class problem.
- Comparing the performance of non-sequential models, sequential models and sequential models with a decay factor versus our novel approach to show the importance of good encoding of the hidden relationship between diagnoses and the importance of considering the time factor.
- Integrating an explainable component so that physicians have greater confidence in the decision made by the model.

With the proposed approach, we aim to provide an AI model that helps medical professionals to overcome the challenge of low diagnostic accuracy. To improve the diagnostic process, we not only propose a deep learning approach with sufficient accuracy, but we also ensure that our algorithm is trained on a sufficiently large quantity of real-world data that is available during the clinical process and propose an application scenario for our algorithm.

The remainder of this paper is organized as follows: In Section 2, we present background works. In Section 3, we formally represent the dataset

and describe our model. Section 4 is devoted to empirically evaluating our model against our competitors on different metrics. We have also carried out various ablation studies to show how the model works in different configurations. In Section 6, we discuss the explainability of the results. In Section 7, we discuss possible use case scenarios, limitations and further research. Section 8 recalls the paper's main points and contributions and outlines future work.

2. Related Work

Data science aims to develop computational models that can automatically infer hidden patterns from data to predict results. Predictions can be based on single or multi-modal data sources [35]. For depression detection, several data sources like audio [36, 37, 38], EEGs [39, 40, 41, 42], IoT or wearable data [43, 44, 45], medical images [46, 47] and text data [48, 49, 50, 51] have been investigated. However, these data must be specifically collected and available for a decision support system. We argue that data generated during the clinical process (like diagnosis data) have a much higher chance to power DDSSs because data availability and quality are lower and privacy issues are fewer. Examples of these data include medical claims or electronic health record data. Promising results on using medical claims data for calculating the risk of suicide prevention have been reported in [25]. Medical claims data are also used to predict reactions to antidepressant treatment [52]. However, studies using machine learning [33] or rule-based approaches [53] on medical claims data for depression screening still report low accuracy metrics.

With a growing volume of data and features and increased computing power, deep learning starts to outperform traditional ML methods [54]. Traditional ML methods typically require good feature selection and a significant amount of feature engineering to ensure that the features used comply with the model's assumptions. On the other hand, deep learning uses a large, multi-layer network structure, allowing it to take raw input features and still be able to learn hidden patterns in data. Deep learning architectures can be distinguished by the structure determining how the network's artificial neurons are connected. For processing sequential and/or structured/unstructured data (like historical diagnoses and medication, clinical notes and images), *recurrent neural networks* (RNNs) [55], *convolutional neural networks* (CNNs) [56], *transformers* [23] and *graph neural networks*

(GNNs) [57] are perfect candidates. Due to their high performance on non-medical tasks addressed with the same data structure as medical applications, their utilization in the medical field has increased significantly. For example, in [58], GNNs are combined with a pre-trained transformer-based model, namely BERT (*Bidirectional Encoder Representations from Transformers*) [59] for medical code representation and medication recommendation. Also, in [60], a pre-trained BERT (specifically its transformer-encoder component) is used to predict 30-day hospital readmissions from clinical notes. In [61], a cost-sensitive formulation of *long short-term memory networks* (LSTM) [62] is proposed to predict 30-day readmission of congestive heart failure patients. Similar work using machine learning and deep learning approaches to predict mortality and readmission of in-hospital cardiac arrest patients with EHR was also conducted in [63]. In [64], a convolutional graph transformer is developed to learn the hidden structure of Electronic Health Record (EHR) data for graph reconstruction while predicting hospital readmission. While some minor modifications had to be made to the core of the aforementioned deep learning models to deal with medical data efficiently, even more significant changes were needed to deal with the ubiquitous irregular time series in the medical field. For example, several studies [30, 28, 65] have focused on redesigning RNNs to better handle irregular physiological time series data and thus improve the accuracy of downstream medical tasks.

Regarding our main concern, namely the detection of depression, we note several studies based on deep learning methods, such as [66, 67, 68], which have used social network data rather than medical claims data. For example, in [69], an LSTM-based model is coupled with an attention mechanism to detect depression from users' tweets. The dataset was balanced (oversampling or undersampling) to address the imbalanced class problem. In [70], a data augmentation framework based on topic modelling is proposed to solve the problem of imbalanced classes when detecting depression. In contrast to approaches based on social network data, [70] uses patients' responses recorded during encounters with doctors. Unlike the technique we use (the cost-sensitive loss function), the undersampling or oversampling technique has the disadvantage of altering the natural distribution of the data used in the study. In addition, they may lose some information or add noise. Although several depression detection studies have been conducted using social network data, some, like ours, have used medical claims data [71, 72]. In [32], a bidirectional deep learning model is proposed—a pre-trained and fine-tuned version of the BERT model. Compared to our proposal, where only

diagnoses and patient demographics are used, the approach in [32] uses additional modalities such as procedures, medications and clinical notes. Our approach, as in [32], relies on the self-attention component to quantitatively assess the association between clinical codes; however, it does not consider the time interval when modelling consecutive visits.

As we can see, none of the aforementioned models addressing the problem of depression detection have combined *self-attention* with GRU-decay for better data representation and efficient integration of temporal information, respectively. Additionally, unlike some, we use real and voluminous medical datasets and do not apply any undersampling or oversampling that may remove relevant information or introduce noise. Instead, we use a cost-sensitive loss function to deal with the problem of imbalanced classes.

3. Method

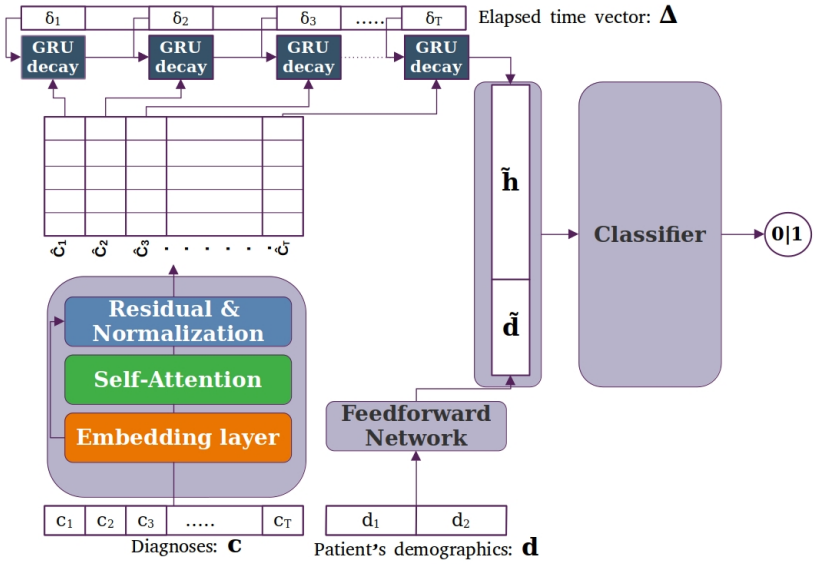
Detecting depression from claims data involves considering three important aspects: learning the hidden relationships between diagnoses; filtering out the irrelevant ones; and relying more on recent diagnoses. In addition, we may face an imbalanced class problem, as there are naturally fewer sick patients than healthy ones.

In this section, after introducing data notation, we formally describe how a self-attention layer is stacked with GRU-decay to cover the aforementioned aspects. Self-Attention is dedicated to learning hidden relationships across diagnoses and filtering out the irrelevant ones. At the same time, GRU-Decay allows detection based on the most recent diagnoses. We name this combination Att-GRU-decay. The output of Att-GRU-decay, which is the global health status of the patient, is combined with the patient's demographics and fed into a classifier that we also describe formally in the following subsections. Finally, we present the loss function used to address the imbalanced class problem. The overall model architecture is depicted in Fig. 1.

3.1. Data Notation

Let $\mathcal{D} = \{\mathbf{c}_n, \mathbf{d}_n, \Delta_n, y_n\}_{n=1,2,\cdots,N}$ where $\mathbf{c}_n = [c_1, c_2, \cdots, c_{t=T}]$ is the set of diagnoses (ICD-10 codes) of the patient n recorded at date index $t = 1, 2, \cdots, T$. If a patient suffers from depression, the highest date index T is the one preceding the date on which the depression was detected in the patient. Otherwise, the highest date index T is that of the last diagnosis made. $\mathbf{d}_n = [d_1, d_2]$ is the patient's demographic vector. d_1 is the age and

Figure 1: Att-GRU-decay architecture. The embedding layer encodes diagnoses into continuous vectors; Self-attention learns the hidden relationship between the diagnoses pair; and the Residual & Normalization retain the initial information and prevent gradient problems. GRU-decay learns the sequential pattern of diagnoses while taking into account the elapsed time between visits δ_t and generates a context vector $\tilde{h}$, which is a latent representation of the patient's health status. $\tilde{h}$ is concatenated with the latent representation of the patient's demographics and passed through the classifier.



d_2 is the gender. $\Delta_n = [\delta_1, \delta_2, \cdots, \delta_{t=T}]$ is the elapsed time vector. More precisely, δ_t with $t > 1$ and $t < T$ is the time difference between the recorded date of the medical code c_t and c_{t-1}. $\delta_1 = \delta_T = 1$. y_n is the depression state of the patient n: equal to 0 if the patient has never suffered from depression; otherwise, it is equal to 1.

3.2. Self-Attention

This section presents how diagnoses are transformed into a vector embedding and passed through a self-attention layer responsible for learning hidden relationships between diagnoses and filtering out the ones irrelevant to the downstream task.

Since neural networks require real numbers as inputs, the first step is to map each diagnosis to a vector of real numbers. For that, we use an *embedding layer*, which, based on the co-occurrence of diagnoses, will associate with each diagnosis c_t[7] a vector embedding $\tilde{\mathbf{c}}_t$ obtained as part of a matrix $\tilde{C}$ of all vector embeddings as follows:

$$\tilde{C} = Embedding_\theta(\mathbf{c}) \tag{1}$$

where θ are the learnable parameters of the *embedding layer*. $\tilde{C} \in \mathbb{R}^{T \times l}$ is the matrix of diagnoses encoded, where each row t of $\tilde{C}$ is the vector embedding $\tilde{\mathbf{c}}_t$ of the diagnosis c_t, and l is the dimension of the embedding space.

To discover the latent relationships between diagnoses and filter out diagnoses that might not be relevant for detecting whether a patient suffers from depression, we pass $\tilde{C}$ through a self-attention layer. Throughout self-attention computations, we calculate an attention filter from a query matrix Q and a key matrix K to encode hidden relationships between diagnoses. This attention filter is then multiplied by a value matrix V to obtain a filtered version $\tilde{C}'$ of $\tilde{C}$. In other words, those vector embeddings $\tilde{\mathbf{c}}_t$ that will be detected as irrelevant for determining whether a patient suffers from depression will have coefficient values close to zero. The formula for calculating $\tilde{C}'$ is

$$\tilde{C}' = \underbrace{Softmax\left(\frac{QK^\intercal}{\sqrt{d_k}}\right)}_{attention\ filter} V \tag{2}$$

where $Q = \tilde{C}W_Q \in \mathbb{R}^{T \times j}$, $K = \tilde{C}W_K \in \mathbb{R}^{T \times j}$ and $V = \tilde{C}W_V \in \mathbb{R}^{T \times j}$ are three different linear transformations of $\tilde{C}$. W_Q, W_K and W_V are learnable parameters, $j = l$ is the dimension of each linear space, d_k is the dimension of the key vectors, and as usual, $K^\intercal$ stands for the transpose of K. $\tilde{C}'$ is normalized to prevent exploding values. The normalized version of $\tilde{C}'$ is then added to $\tilde{C}$ to preserve initial relevant information that might be lost during self-attention computation and to prevent gradient problems. Residual is the sum of an input x with the output $y = f(x)$ [73]. The final output of the self-attention layer is then equal to:

[7] As the following formulas are valid for all patients, we omit the subscript n in the sequel.

$$\hat{C} = \underbrace{\tilde{C} + Normalize(\tilde{C}')}_{residual} \tag{3}$$

$\hat{C} \in \mathbb{R}^{T \times j}$ is a matrix, where each row t is the final embedding representation $\hat{\mathbf{c}}_t$ of a corresponding diagnosis c_t.

As some diagnoses may have been made long ago, assessing their significance in terms of when they were made is crucial. Thus, in Section 3.3, we formally show how we apply a decay factor on the hidden layer of GRU so that past diagnoses cannot have the same level of importance as recent ones.

It is worth mentioning that self-attention aims to encode the hidden correlation between pairs of clinical codes, while GRU-decay encodes the sequential order of visits, taking into account the time elapsed between them. We could have used the positional encoding technique implemented in the original Transformer [23] to model the sequential order of visits. However, since the positional encoding vectors are static, we would not have been able to capture the variation in elapsed time between successive visits effectively.

3.3. GRU-Decay

Although some patients may suffer from depression without prior symptoms, we can detect those who do with specific earlier symptoms. This then requires browsing the patient's historical diagnoses. Since diagnoses are described by a set of clinical codes recorded over time, depression detection can be approached as both time-series forecasting and NLP tasks.

With RNNs and their variants having shown spectacular results on time series forecasting and NLP tasks, e.g. [74, 75], we can use them to model the patient's status while considering the time at which each diagnosis has been recorded. Since RNNs suffer from gradient problems when processing long sequences, we use their variant Gated Recurrent Unit (GRU), which addresses this problem. GRU is mathematically defined as follows:

$$\mathbf{z}_t = \sigma_g(\hat{\mathbf{c}}_t W_z + \mathbf{h}_{t-1} U_z + b_z) \tag{4}$$
$$\mathbf{r}_t = \sigma_g(\hat{\mathbf{c}}_t W_r + \mathbf{h}_{t-1} U_r + b_r) \tag{5}$$
$$\bar{\mathbf{h}}_t = \phi_h(\hat{\mathbf{c}}_t W_h + (\mathbf{r}_t \odot \mathbf{h}_{t-1}) U_h + b_h) \tag{6}$$
$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \bar{\mathbf{h}}_t \tag{7}$$

where $\mathbf{h}_{t-1}$ with $(t-1) \geq 0$ is the hidden state of the medical code embedding $\hat{\mathbf{c}}_{t-1}$; $\mathbf{z}_t$ and $\mathbf{r}_t$ are the update and reset gates associated with the

medical code embedding $\hat{\mathbf{c}}_t$, respectively; $\bar{\mathbf{h}}_t$ is the hidden intermediate state; $\mathbf{h}_t$ is the hidden state of the current input $\hat{\mathbf{c}}_t$ and also the GRU's output; W_z, W_r, W_h, U_z, U_r, U_h, b_z, b_r; and b_h are GRU training parameters; and $\odot$ denotes the Hadamard product as usual.

The GRU's equations (4)–(7) assume that the elapsed time δt between the recording dates of two consecutive diagnoses is regular. This assumption is not valid since δt may vary. In addition, this variation might be high. It is then crucial to weight each hidden state $\mathbf{h}_t$ of $\hat{\mathbf{c}}_t$ according to δt so that the model places more importance on recent diagnoses than those made a long time ago. Therefore, we introduce a decay factor in the GRU and multiply it by $\mathbf{h}_t$ to obtain a new hidden state $\tilde{\mathbf{h}}_t$. A GRU with a decay factor applied to its hidden state is called GRU-decay. Except for the hidden state, it has the same structure as a GRU. $\tilde{\mathbf{h}}_t$ is obtained as follows:

$$\tilde{\mathbf{h}}_t = exp(-max(0, \delta_t \tilde{W} + \tilde{b})) \odot \mathbf{h}_t \tag{8}$$

where $\tilde{W}$ and $\tilde{b}$ are learnable parameters. $\tilde{\mathbf{h}}_{t=T} = \tilde{\mathbf{h}}$ can be interpreted as a latent summary of the patient's health status.

As patient demographics (gender and age) are important factors to study for depression, we formally describe, in Section 3.4, how they are combined with the latent summary of patient health status $\tilde{\mathbf{h}}$ and then run through the classifier to predict whether the current patient will suffer from depression.

3.4. Depression Detection: Classifier

To calculate the likelihood that a patient will be detected as a depressed patient, we first extract information from patient demographics via a *feed-forward neural network* (FNN) and combine the extracted information with $\hat{\mathbf{h}}_t$. In end-to-end fashion, the result of this combination is fed into a set of stacked FNNs that play the role of the classifier. Formally, the classifier is

$$\hat{y} = f^1_{\beta_1} \circ f^2_{\beta_2} \circ \cdots \circ f^P_{\beta_P}(\langle \tilde{\mathbf{h}}, \tilde{\mathbf{d}} \rangle) \tag{9}$$

$$\tilde{\mathbf{d}} = g_\alpha(\mathbf{d}) \tag{10}$$

where g_α is an FNN with *Relu* as an activation function; α is a set of learnable parameters of g; $\tilde{\mathbf{d}}$ is the latent representation of patient's demographics vector; $f^2_{\beta_2} \circ \cdots \circ f^P_{\beta_P}$ is P stacked FNNs with *Relu* as an activation function; $\beta_2, \cdots, \beta_P$ are learnable parameters; $f^1_{\beta_1}$ is the final FNN with *sigmoid* as an

activation function and β_1 as its learnable parameters; and $\hat{y} \in [0, 1]$ is the likelihood that a patient suffers from depression.

We used a weighted binary cross-entropy as a loss function [76] to adjust the models' parameters while dealing with the problem of imbalanced classes. It is defined as follows:

$$\mathcal{L}_{wbc} = -\frac{1}{N}\sum_{n=1}^{N}\Big(w_1 * y_n * \ln(\hat{y}_n) + w_0 * (1 - y_n) * \ln(1 - \hat{y}_n)\Big) \tag{11}$$

where $w_0 = 1$ and $w_1 = N_0/N_1$ are the weighted factors of class 0 and 1, respectively. w_1 allows penalizing the model more when the class 1 is misclassified. Indeed, this choice is justified because we are dealing with imbalanced classes, i.e. the number of patients suffering from depression is much lower than those who do not suffer from it.

4. Experimentation

4.1. Settings

We coded the proposed model using Python 3.0 and the machine learning libraries Keras 2.4.3 and TensorFlow 2.4.0. All remaining pre-processing and performance evaluation was done with the libraries NumPy, Pandas and Scikit-learn. Finally, we ran the code on a cluster node with the following characteristics: An AMD Threadripper 3960X processor with 24 cores and 48 threads, 128 GB of memory, and an NVidia 3090 GPU with 24 GB of graphics memory.

4.2. Data

Our dataset was queried from the EHIF data warehouse and includes information on gender, birth year, ICD-10 coded primary and secondary diagnoses and the date of the treatment bill (diagnosis date) from $812,853$ people (15 years or above) with a total of $26,973,943$ diagnoses between 2018 and 2022. The data consist of all publicly insured people in Estonia with a depression diagnosis[8] ($80,243$ patients with $4,252,213$ diagnoses). The control group consists of $732,610$ patients (with $22,721,730$ diagnoses), of

[8]To overcome potential data leakage, we considered all diagnoses starting with F32 (major depressive disorder) and F33 (recurrent depressive disorder) as 'depression'.

which 498, 764 people (with 10, 779, 835 diagnoses) did not have a psychiatric disorder diagnosed and 233, 846 patients (with 11, 941, 895 diagnoses) had a psychiatric disorder other than depression. The percentage of insured people in Estonia is above 93.63% [77], so we are confident that our dataset is representative of the entire population.

Diagnoses were coded based on ICD-10. Each ICD-10 code consists of an alpha character known as a chapter, two digits describing the disease category, a dot and additional digits representing more details like the cause, location, severity or other clinical information (sub-categories). For example, F32.2 is the code for major depressive disorder, single episode, severe without psychotic features. F stands for mental and behavioral disorders; F30–F39 are codes for mood [affective] disorders; and F32 is the category of major depressive disorder, single episode. The '.2' at the end of F32.2 specifies the severity. All patients with the ICD-10 codes F33.x or F32.x are classified as patients with depression. For the latter, only diagnoses made before being diagnosed with depression are taken into account in the study. Those which followed the depression diagnosis are ignored. Figure 2 shows the data extraction process.

The Research Ethics Committee of the National Institute for Health Development (TAIEK[9]) approved this study's research design and data usage (Decision No. 1148).

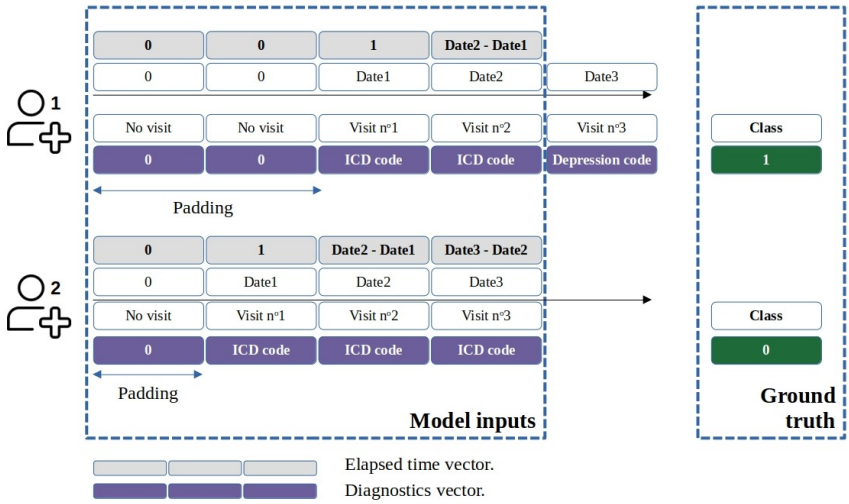
4.3. Model and Training Hyperparameters

We performed an extensive grid search over embedding_space= $\{50, 80, 100\}$, dimenssion_linear_space= $\{32, 64, 80, 100\}$, GRU_decay_units= $\{30, 50, 80, 100\}$, demographics_FNN_units= $\{10, 20, 30\}$, classifier_FNN_units= $\{20, 30, 50\}$, number_epochs= $\{20, 30, 40, 50, 60, 70, 80\}$, optimizer= $\{Adam, SGD, RMSprop\}$ to find the optimal value for each hyperparameter of the model. The values retained for each hyperparameter are as follows:

- The dimension of the embedding space was set to 50.
- We used a mask on the embedding layer to skip padding values during the calculation.
- For the attention layer, we set the dimension of linear spaces at 80.

[9]Tervise Arengu Instituudi inimuuringute eetikakomitee

Figure 2: Data extraction process from two patients. For padding values, we assign 0 as elapsed time. For the first diagnosis (ICD code of the first visit), we assign 1 as the elapsed time. When a depression code is observed in the diagnosis list, the associated ground truth is 1. All diagnoses after the first depression diagnosis are ignored. On the other hand, when no depression code is observed, the ground truth associated with the sample is 0.



- The number of GRU-decay units was set to 50.
- We applied a dropout of 0.5 on the hidden layer of GRU-decay to prevent gradient problems.
- Concerning the FNN dedicated to the extraction of demographic features of patients, we defined the number of units as 10.
- The classifier comprises two stacked FNNs, each with 20 and 1 units, respectively.

Once more, using the grid search technique, we defined the training hyperparameters as follows:

- The number of epochs was set to 20.
- The batch size was set to 1500.

- We used *Adam* as the optimizer.
- The learning rate was set to 0.001.

Table 1 summarizes all the hyperparameter values.

Table 1: Model and training hyperparameters

Hyperparameters	Values
Dimension of the Embedding space	50
Dimension of Linear spaces	80
Number of GRU-decay units	50
GRU-decay dropout	0.5
Number of FNN units of patient's demographics	0.5
Number of FNN units of the classifier	20 & 1
Number of epochs	20
Batch size value	1500
Optimizer	*Adam*

4.4. Results

To assess the performance of our model, we use the *area under the ROC Curve* (AUC) and the *area under the precision-recall curve* (AUPRC) as metrics. The *precision-recall* curve is a function of *recall* (12) on the x-axis and *precision* (13) on the y-axis. The *receiver operating characteristic* (ROC) curve is a function of *false positive rate* (14) on the x-axis and *recall* on the y-axis.

$$Recall = Sensitivity = \frac{|TP|}{|TP| + |FN|} \tag{12}$$

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{13}$$

$$False\ Positive\ Rate = \frac{|FP|}{|FP| + |TN|} \tag{14}$$

where $|TP|$ is the number of true positives, $|FN|$ the number of false negatives, $|TN|$ the number of true negatives and $|FP|$ the number of false positives. Indeed, by varying the threshold when calculating recall, precision and the false positive rate, these metrics avoid biased scores caused by

Table 2: AUC and AUPRC scores on depression detection task over 5-cross validation. ± denotes the standard deviation

Models	AUC	AUPRC
LR	0.813 ± 0.002	0.296 ± 0.003
CNN-LSTM	0.849 ± 0.002	0.394 ± 0.009
LSTM	0.848 ± 0.001	0.385 ± 0.005
FNN	0.837 ± 0.002	0.374 ± 0.006
GRU-decay	0.989 ± 0.001	0.972 ± 0.001
GRU-Δt	0.986 ± 0.002	0.961 ± 0.005
Att-GRU-decay	**0.990 ± 0.001**	**0.974 ± 0.002**

the high number of non-target classes, i.e. the class 0. They are suitable for assessing the performance of models in the face of an imbalanced class problem.

We compare the average AUC and AUPRC scores obtained over 5-fold cross-validation with those of the following models: logistic regression (LR); feedforward neural network (FNN); long short-term memory (LSTM); convolutional neural network combined with LSTM (CNN-LSTM); gated recurrent unit with a decay factor (GRU-decay); and a gated recurrent unit taking as inputs diagnostic vectors concatenated to the elapsed time vectors (GRU-Δt). All results are reported in Table 2.

From Table 2, we can clearly see that our proposed model achieves the best performances. Although the GRU-decay and GRU-Δt results are very accurate, ours are slightly better. Compared to our Att-GRU-decay model and the GRU-decay model, GRU-Δt is less accurate because it does not incorporate any explicit techniques to better learn existing patterns between diagnoses and time. The slight superiority of our model highlights the additional contribution of the self-attention layer in the decision-making process. Indeed, unlike the GRU-decay, which only benefits from decay factors that prevent the prediction from being based on diagnoses made a long time ago, our model, thanks to the self-attention mechanism, will also detect hidden patterns existing between diagnoses that may be the cause of possible depression in the patient. Where the difference between the AUC scores of the models is not so large, the AUPRC scores of our model and the GRU-decay model far exceed those of the other competitors. This huge difference reveals how crucial it is to weigh the significance of the diagnoses according to their respective recording dates.

It is not surprising that the LR model, which is a traditional machine learning model, performs worse than the other models, which are deep learning models. Indeed, unlike machine learning, which is somewhat dependent on feature engineering, deep learning can extract hidden features by itself thanks to its non-linear functions and therefore does not need feature engineering. This property makes deep learning models more accurate than machine learning models when processing data with complex patterns. Machine learning models can sometimes achieve results similar to or better than deep learning models [78]. Moreover, they are more explainable. Another aspect that reveals the results in Table 2 is the low accuracy of non-sequential models such as LR and FNN compared to others designed for sequence modelling. Thus, we conclude that processing patients' diagnoses at different dates with non-sequential models leads to losing temporal patterns in depression detection. Compared to GRU-decay and our model, CNN-LSTM and LSTM, also models designed to handle sequential data, failed because they processed diagnoses as if they were made at regular time intervals.

As AUC and AUPRC are calculated from different thresholds, we also investigate the specificity (15) and the sensitivity of the models on fixed threshold values of 0.5 and 0.8. This second evaluation was carried out on a single loop of the 5-cross validation.

$$Specificity = \frac{|TN|}{|TN| + |FP|} \tag{15}$$

Indeed, the higher the threshold, the more confidence practitioners have in the model's outcome. A higher threshold is even more important in the medical field, as misdiagnoses can have irreversible consequences. The specificity and sensitivity scores of all models calculated from the confusion matrices in Fig. 3 are reported in Table 3. We also report the training and testing time for each model to give an idea of how long it will take for each of them to produce results in a real deployment. The ROC curves and precision-recall curves of the evaluated models are shown in Figure 4.

Table 3 and Figure 4 show that GRU-decay, GRU-Δt and ours obtain the best specificity scores, sensitivity scores, ROC curves and Precision_Recall curves. These scores again show how incorporating a decay factor to handle better diagnoses recorded at different dates improves the classification task. We note that almost all models provide satisfactory results with a threshold set to 0.5. We assume that these results are due to the large qualitative amount of data and the weighted binary cross-entropy, which improves the

Figure 3: Confusion matrices

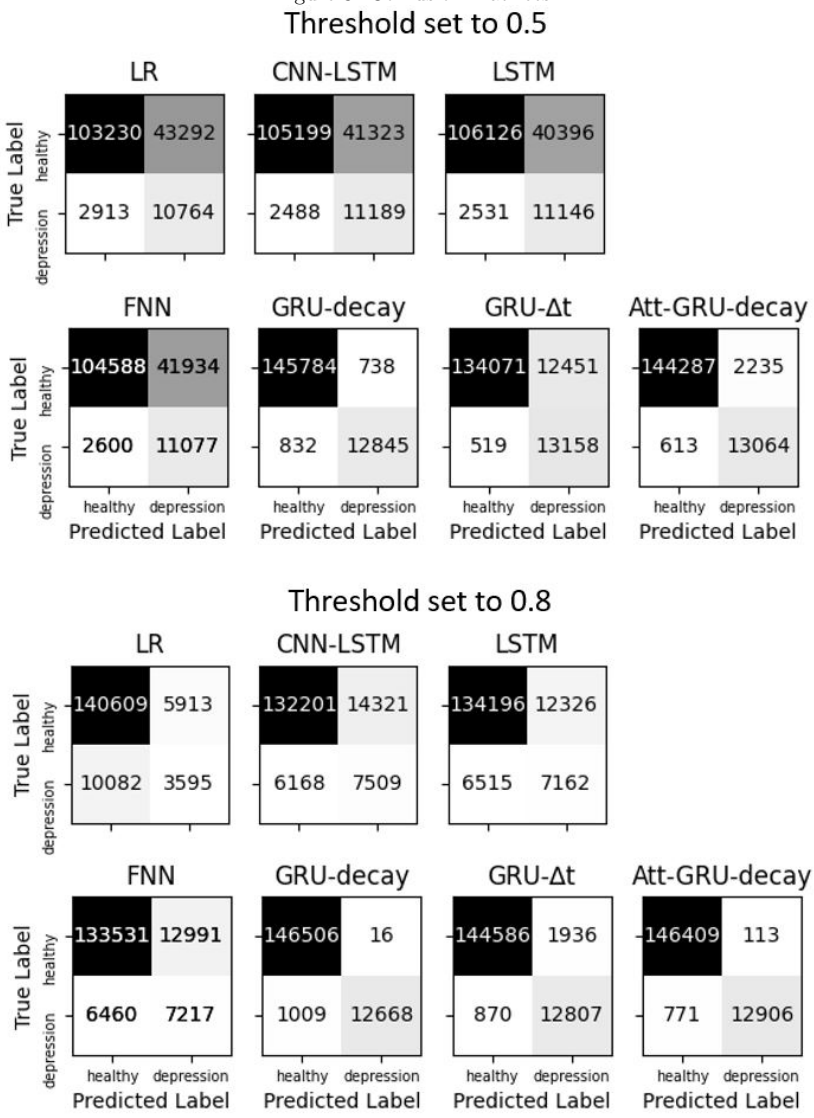
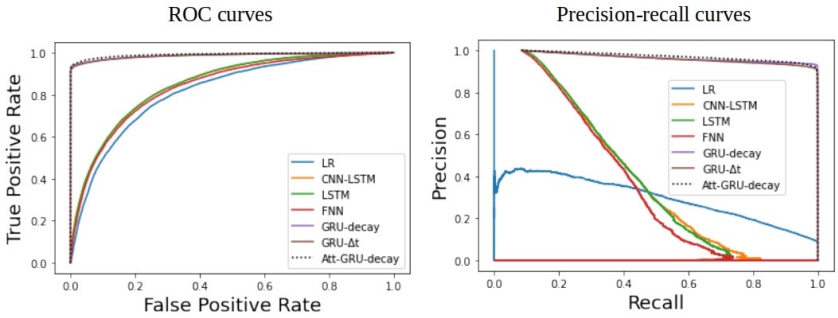


Table 3: Specificity and sensitivity scores on the depression detection task

	Specificity		Sensitivity			
	Threshold					
	0.5	0.8	0.5	0.8	Time (min)	
Models					Train	Test
LR	0.705	0.960	0.787	0.263	8.680	**0.008**
CNN-LSTM	0.718	0.902	0.818	0.549	8.299	0.035
LSTM	0.724	0.916	0.814	0.523	30.513	0.050
FNN	0.714	0.911	0.810	0.528	**1.312**	0.006
GRU-decay	**0.995**	**0.999**	0.939	0.926	49.547	0.068
GRU-Δt	0.962	0.987	0.962	0.936	2.877	0.102
Att-GRU-decay	0.985	**0.999**	**0.955**	**0.944**	56.754	0.102

Figure 4: ROC and Precision-recall curves



models' ability to classify the minority class, i.e. the depressed patient. If a threshold is set to 0.5, the sensitivity scores for all models are fairly accurate. We find a considerable drop in the performance of the LR, CNN-LSTM, LSTM and FNN models when the threshold is set to 0.8. On the other hand, the GRU-decay model and ours remain very accurate. Although the specificity score of the GRU-decay model with a threshold of 0.5 is better than that obtained with ours, with the other configurations, our model is better overall. It is worth mentioning that, despite the high threshold value of 0.8, we obtained spectacular sensitivity and specificity scores close to 1. In verbal form, among the 146,522 non-depressed patients in the training set, our model correctly classifies 146,409 with a probability of 0.8%. For

the 13,677 depressed patients, our model correctly classifies 12,906 with a probability of 0.8%.

For classification problems such as those related to medicine, the output of the models must be very accurate to avoid misdiagnoses leading to inappropriate treatment. Especially for the early detection of psychiatric diseases such as depression, the model's sensitivity is crucial. With the quantitative results we have obtained, we are very confident that our model can help medical professionals in their decision-making to detect patients with depression faster and thus significantly reduce the misdiagnosis rate.

We note that in terms of training and testing times, our model takes the longest. This is partly due to the number of parameters (97,121) and the time complexity of the self-attention and GRU-decay mechanisms. Despite having the longest test duration, 0.102 minutes is still sufficient for using it in the clinical process. The number of parameters in the competing models is shown in Appendix A.6.

In the next section, we conduct different ablation studies to show how the model works in different configurations.

5. Ablation studies

We have devoted this section to evaluating the model in the following configuration: i) without the decay factor; and ii) with and without patient demographics.

Without the decay factor. In Table 4, we observed a considerable drop in performance when the decay factor is not taken into account. These results support our assertion regarding the importance of accounting for irregular elapsed time between visits. Indeed, the normal GRU fails because it processes diagnoses as if they were made at regular intervals and is therefore unable to capture the correct underlying temporal pattern of diseases.

Table 4: Evaluation of the model without the decay factor over 5-cross validation

Models	AUC	AUPRC
Att-GRU	$\mathbf{0.853 \pm 0.001}$	$\mathbf{0.405 \pm 0.007}$
Att-GRU-decay	$\mathbf{0.990 \pm 0.001}$	$\mathbf{0.974 \pm 0.002}$

With and without patient demographics. The AUC and AUPRC scores in Table 5 show that patient demographics have little influence on the detection of depression. We can see that without patient demographics, the performance of the model is not affected. However, when patient demographics are used exclusively, model performance drops significantly. We conclude that the model can still produce accurate results when patient demographics are not available. Scientific literature suggests an impact of demographic factors like

Table 5: Evaluation of the model without patient demographics and exclusively with over 5-cross validation. ex/pd stands for exclusively with patient demographics, and wo/pd stands for without patient demographics

Models	AUC	AUPRC
Att-GRU-decay ex/pd	**0.647 ± 0.002**	**0.131 ± 0.002**
Att-GRU-decay wo/pd	**0.990 ± 0.001**	**0.972 ± 0.001**
Att-GRU-decay	**0.990 ± 0.001**	**0.974 ± 0.002**

gender [79] or age [80] on the likelihood of getting depression. We assume that this effect is not visible in our ablation study because the model learns gender and age trends through associated diseases.

As quantitative results are not sufficient to guarantee the veracity of a model in medical applications, we also propose a component for extracting disease patterns that influence the model output to be able to give a qualitative interpretation of the model behavior (see Section 6).

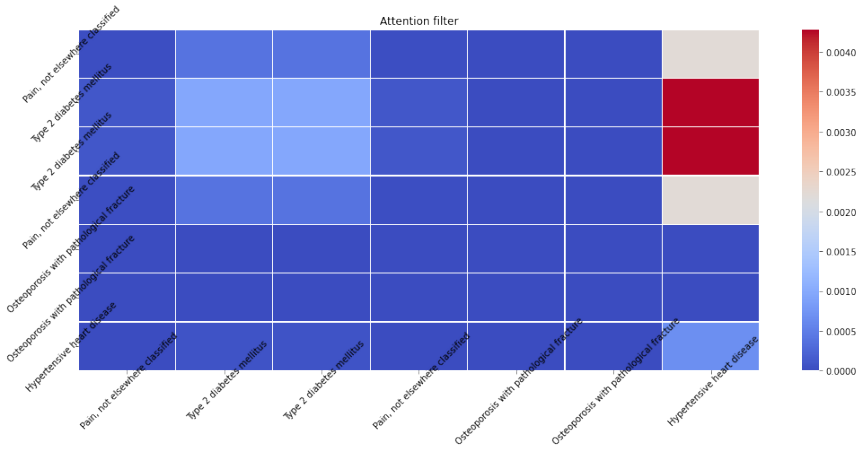
6. Uncovering Disease Patterns

The following section shows the interaction between features in the attention layer of our model.

Apart from the benefits of increased prediction accuracy, we use self-attention to provide insights into the disease relationships the model has learned. We propose using this to give medical professionals a better understanding of the model by showing that it can correctly identify commonly known disease correlations. Those disease correlations can also be used to infer rules and find indicator diseases [53].

The alignment matrix in Fig. 5 shows a given patient's last seven ICD-10 codes on the x-axis and how our trained neural network associates them with each other. The color indicates the strength of the correlation, from blue (not correlated) to red (strongly correlated). In this example, our trained network

Figure 5: Attention filter – Example 1



identified a strong correlation between heart failure and type 2 diabetes. This correlation is already well known in medicine and shows how the Att-GRUD-decay could infer it from the training data.

Now, consider the second patient (Fig. 6). We see the last ten diagnoses, from which the model identified that oesophagitis is correlated with migraine, dorsalgia and abdominal and pelvic pain. While abdominal and pelvic pain could logically make sense, there is currently no strong medical evidence for a correlation with migraine or dorsalgia. Nevertheless, some forms of migraine trigger strong nausea, which could lead to oesophagitis and spinal problems, manifesting as dorsalgia and negatively influencing migraines.

The third patient (Fig. 7) shows a strong correlation between the need for immunization against other single viral diseases and sleep disorders and retinal disorders. We are unaware of any medical evidence of a correlation between those ICD-10 codes.

This example demonstrates how the model learned and can find reasonable connections from large data sets. Still, not all correlations are evidence-based from a medical perspective.

Generally, the correlations found can be sorted into three categories:

1. *True correlations*, which are (based on our current medical knowledge) reasonable (Fig. 5, and potentially Fig. 6).

Figure 6: Attention filter – Example 2

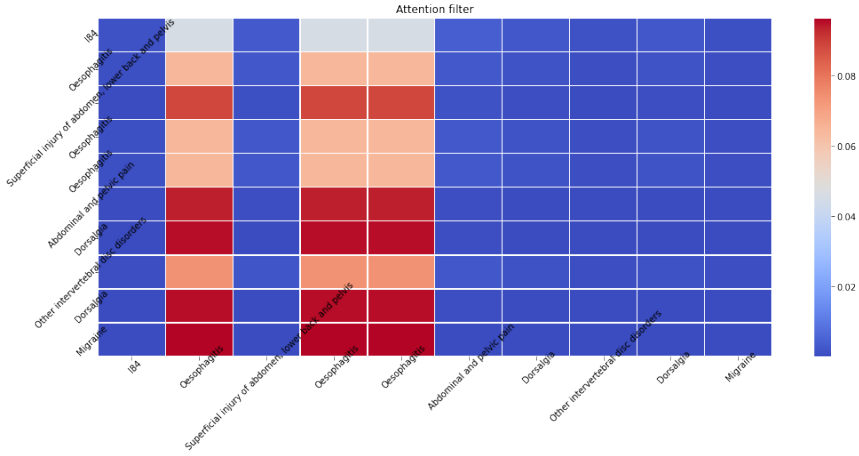
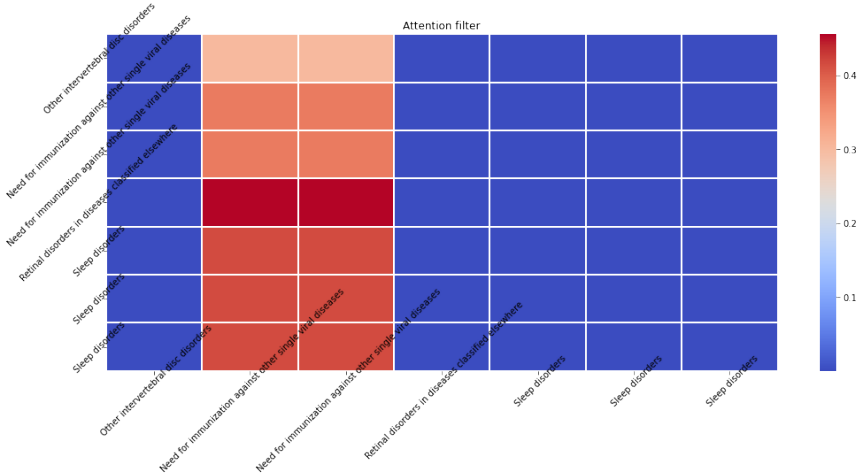


Figure 7: Attention filter – Example 3



2. *"Hallucinations"* of the deep learning network, i.e. output that does not seem to be justifiable based on the training data (potentially Fig. 6 and Fig. 7).
3. *Potentially true correlations*, which we currently cannot grasp because they exceed today's medical knowledge (Fig. 6, potentially Fig. 7).

So, while the prediction accuracy of our model is high, the individual correlations shown by the self-attention still need to be evaluated carefully.

It is important to note that these correlations are only based on the attention layer of our model. They do not offer explainability of other parts, e.g. the GRU component, of our model.

7. Discussion

Several high-performing AI models have already been proposed in the healthcare sector. Still, success stories of AI providing real clinical value are rare. The reasons for this include a lack of data availability, integration into clinical processes and lack of trust due to the black-box characteristics of the models. In the previous sections, we demonstrated that our novel Att-GRU-decay model outperforms the current state of the art. In this section, we elaborate on a possible application scenario to demonstrate how this model could improve the status quo while avoiding the above-mentioned pitfalls.

Since one of the main problems in psychiatry is that patients with psychiatric disorders are often diagnosed late, we propose to use this model to screen patients when they visit a healthcare professional proactively. This makes sense, especially for general practitioners (GPs) with high patient turnover. The model can be plugged into the GP's systems and rolled out at the insurance provider level or on a national level on top of an NHIS via a RESTful [81] API. If the GP enters the diagnosis at the end of the visit, our model enables the doctor's IT systems to send an alert if the patient is thought to have undiagnosed depression. The GP can then re-evaluate the decision, using our explainability component, and refer the patient to a specialist for further treatment in the case of a true positive prediction. Since the proposed system operates on diagnoses from medical claims data, which medical professionals capture during their work anyway, no additional effort is needed. This allows seamless integration into the current clinical workflow. Because of the high specificity, we assume the risk of alert fatigue is low. On the other hand, if we compare our sensitivity of 94.4% to

the reported 50.1% (95% CI: 41.3 to 59.0) sensitivity of GPs for diagnosing depression [4], we see that our model has the potential to decrease the number of undiagnosed depression patients significantly. It even outperforms population-level screening questionnaires, such as the PHQ-9, which has a sensitivity of 88% and a specificity of 88% [82]. In addition, time-wise, the suggested approach outperforms the current use of questionnaires and assessment scales. Current depression assessment instruments, such as the Beck Depression Inventory [9], the Hamilton Depression Rating Scale [10] or the Montgomery-Åsberg Depression Rating Scale [83], take between 15 and 30 minutes to complete. At the same time, our proposed screening approach outputs results in seconds.

This use case can be expanded to screening other diseases in domains other than mental health as long as the data utilized have the same structure. Since we operate based on medical claims data available in most countries and cover a wide range of medical information, it should be fairly easy to retrain our model to predict other diseases.

It is important to stress that we are not proposing to replace medical doctors with AI algorithms. We suggest that AI algorithms can be used as screening instruments, assisting doctors by discovering hidden patterns in large volumes of medical data to help them diagnose faster and more accurately. The output of an AI model still needs to be validated, checked against the current patient situation, and communicated. Furthermore, the subsequent steps, i.e. further diagnostic procedures and treatment decisions, still need to be taken by doctors.

We are fairly confident that the model will perform well in a production setting because of the large amount of real-world data used for training and evaluation, which includes nearly every adult Estonian. For further research, the model needs to be evaluated in a randomized control trial (RCT) to obtain further evidence on its usefulness in a clinical setting. One limitation of our study is that the data we used as ground truth might be biased, for instance, because of the previously described low accuracy of human diagnoses, but also because medical claims data are used for billing purposes, which creates an incentive for medical professionals to adapt codes to maximize revenue. An RCT can help show the impact of this potential bias on the usefulness of our proposed model. Another limitation of our research is that we did not use any prescription, laboratory, genomics data or other unobtrusive data sources. We focused solely on diagnostic and socio-demographic data because this is easily accessible during the clinical process without the

need for any specific data collection by the physician or patient. Because of the good results of our approach, we saw the exploration of other data sources as out of scope. Nevertheless, we encourage further research to analyse whether other AI algorithms based on other clinical data sources can give similar or better results. Additionally, we encourage further research to evaluate the described scenario with other digital health evaluation methods to assess usability and efficacy.

Further research is also planned to investigate how well the model can be applied to different diseases using the same kind of data. We see the use of self-attention rather than multi-head attention as a potential limitation in terms of explainability and disease correlations. The use of multi-head attention could potentially find more and deeper hidden disease patterns.

8. Conclusion

In this research, we used the medical claims data of $812,853$ patients with $26,973,943$ diagnoses to evaluate deep learning for depression detection. We contribute by evaluating the most common deep learning algorithms and introducing our novel Att-GRU-decay model, which outperforms other state-of-the-art deep learning models with an AUC of 0.99 and an AUPRC of 0.974. We further describe a potential application scenario for using the proposed model for screening patients in a GP setting. Since the use of real-world data covers nearly every adult Estonian, the excellent accuracy results of Att-GRU-decay, in addition to the proposed use-case scenario with a potential increase in the specificity of depression diagnosis by GPs from 50.1% to as much as 94.4%, we see this research as a potential game changer for psychiatric screening.

Acknowledgements

We want to thank the EHIF for providing access to the data needed for this research. The Estonian Human Research Ethics Committee (TAIEK) of the Institute for Health Development (Decision No. 1148) approved this study's research design and data usage.

Competing Interests Statement

The authors declare no conflicts of interest.

References

[1] H.-U. Wittchen, F. Jacobi, J. Rehm, A. Gustavsson, M. Svensson, B. Jönsson, J. Olesen, C. Allgulander, J. Alonso, C. Faravelli, et al., The size and burden of mental disorders and other disorders of the brain in Europe 2010, European Neuropsychopharmacology 21 (9) (2011) 655–679.

[2] E. A. McGlynn, S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, E. A. Kerr, The quality of health care delivered to adults in the United States, New England Journal of Medicine 348 (26) (2003) 2635–2645.

[3] A. Aboraya, E. Rankin, C. France, A. El-Missiry, C. John, The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis, Psychiatry (Edgmont) 3 (1) (2006) 41.

[4] A. J. Mitchell, A. Vaze, S. Rao, Clinical diagnosis of depression in primary care: a meta-analysis, The Lancet 374 (9690) (2009) 609–619.

[5] S. P. Wamala, J. Lynch, M. Horsten, M. A. Mittleman, K. Schenck-Gustafsson, K. Orth-Gomer, Education and the metabolic syndrome in women., Diabetes Care 22 (12) (1999) 1999–2003.

[6] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, et al., Cost of disorders of the brain in Europe 2010, European Neuropsychopharmacology 21 (10) (2011) 718–779.

[7] S. Z. Williams, G. S. Chung, P. A. Muennig, Undiagnosed depression: A community diagnosis, SSM-Population Health 3 (2017) 633–638.

[8] American Psychology Association, Depression assessment instruments (2023).
URL `https://www.apa.org/depression-guideline/assessment`

[9] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of General Psychiatry 4 (6) (1961) 561–571.

[10] M. Hamilton, A rating scale for depression, Journal of Neurology, Neurosurgery, and Psychiatry 23 (1) (1960) 56.

[11] M. Bertl, P. Ross, D. Draheim, A survey on AI and decision support systems in psychiatry – Uncovering a dilemma, Expert Systems with Applications (2022).

[12] M. Bertl, J. Metsallik, P. Ross, Digital decision support systems for post-traumatic stress disorder – Implementing a novel framework for decision support systems based on a technology-focused, systematic literature review, Frontiers in Psychiatry (2022). doi:10.3389/fpsyt.2022.923613.

[13] M. Bertl, K. J. I. Kankainen, G. Piho, D. Draheim, P. Ross, Evaluation of Data Quality in the Estonia National Health Information System for Digital Decision Support, in: Proceedings of The International Health Data Workshop, 2023.

[14] N. Ivers, A. D. Brown, A. S. Detsky, Lessons From the Canadian Experience With Single-Payer Health Insurance: Just Comfortable Enough With the Status Quo, JAMA Internal Medicine 178 (9) (2018) 1250–1255. doi:10.1001/jamainternmed.2018.3568.
URL https://doi.org/10.1001/jamainternmed.2018.3568

[15] C. Wendt, Changing healthcare system types, Social Policy & Administration 48 (7) (2014) 864–882.

[16] K. Grosios, P. B. Gahan, J. Burbidge, Overview of healthcare in the UK, EPMA Journal 1 (4) (2010) 529–534.

[17] K. McBride, M. Toots, T. Kalvet, R. Krimmer, Leader in e-government, laggard in open data: Exploring the case of Estonia, Revue francaise d'administration publique 3 (2018) 613–625.

[18] S. Lips, V. Tsap, N. Bharosa, R. Krimmer, T. Tammet, D. Draheim, Management of national eID infrastructure as a state-critical asset and public-private partnership: Learning from the case of Estonia, Information System Frontiers (2023). doi:10.1007/s10796-022-10363-5.

[19] J. Metsallik, P. Ross, D. Draheim, G. Piho, Ten years of the e-health system in Estonia, in: CEUR Workshop Proceedings, Vol. 2336, 2018, pp. 6–15.

[20] Estonian National Electoral Committee and the State Electoral Office, Valimised – Voting results in detail (2019).
URL `https://ep2019.valimised.ee/en/voting-result/index.html,Lastaccessedon2022-03-04`

[21] L. Parv, P. Kruus, K. Motte, P. Ross, An evaluation of e-prescribing at a national level, Informatics for Health and Social Care 41 (1) (2016) 78–95.

[22] A. Conneau, H. Schwenk, L. Barrault, Y. Lecun, Very deep convolutional networks for natural language processing, arXiv preprint arXiv:1606.01781 2 (1) (2016).

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[24] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al., Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation, Information Fusion 64 (2020) 149–187.

[25] W. Xu, C. Su, Y. Li, S. Rogers, F. Wang, K. Chen, R. Aseltine, Improving suicide risk prediction via targeted data fusion: Proof of concept using medical claims data, Journal of the American Medical Informatics Association 29 (3) (2022) 500–511.

[26] Z. S. Chen, I. R. Galatzer-Levy, B. Bigio, C. Nasca, Y. Zhang, et al., Modern views of machine learning for precision psychiatry, Patterns 3 (11) (2022) 100602.

[27] Y. Jing, Intelligent assessment of mental health based on multisource information fusion, Journal of Healthcare Engineering 2022 (2022).

[28] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Scientific Reports 8 (1) (2018) 1–12.

[29] Y. Lee, E. Jun, H.-I. Suk, Multi-view integration learning for irregularly-sampled clinical time series, arXiv preprint arXiv:2101.09986 (2021).

[30] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, Y. Li, Brits: Bidirectional recurrent imputation for time series, Advances in Neural Information Processing Systems 31 (2018).

[31] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555 (2014).

[32] Y. Meng, W. Speier, M. K. Ong, C. W. Arnold, Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression, IEEE Journal of Biomedical and Health Informatics 25 (8) (2021) 3121–3129.

[33] M. Bertl, P. Ross, D. Draheim, Predicting psychiatric diseases using autoai: A performance analysis based on health insurance billing data, in: International Conference on Database and Expert Systems Applications, Springer, 2021, pp. 104–111.

[34] B. Hosseinifard, M. H. Moradi, R. Rostami, Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal, Computer Methods and Programs in Biomedicine 109 (3) (2013) 339–345.

[35] S. Thandapani, M. I. Mahaboob, C. Iwendi, D. Selvaraj, A. Dumka, M. Rashid, S. Mohan, Iomt with deep cnn: Ai-based intelligent support system for pandemic diseases, Electronics 12 (2) (2023) 424.

[36] S. Sardari, B. Nakisa, M. N. Rastgoo, P. Eklund, Audio based depression detection using convolutional autoencoder, Expert Systems with Applications 189 (2022) 116076.

[37] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: An efficient deep model for audio based depression classification, in: Proceedings of AVEC'16 – the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 35–42.

[38] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, P. Georgiou, Multimodal and multiresolution depression detection from speech

and facial landmark features, in: Proceedings of AVEC'16 – the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 43–50.

[39] B. Ay, O. Yildirim, M. Talo, U. B. Baloglu, G. Aydin, S. D. Puthankattil, U. R. Acharya, Automated depression detection using deep representation and sequence learning with EEG signals, Journal of Medical Systems 43 (2019) 1–12.

[40] S.-C. Liao, C.-T. Wu, H.-C. Huang, W.-T. Cheng, Y.-H. Liu, Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns, Sensors 17 (6) (2017) 1385.

[41] M. Bachmann, L. Päeske, K. Kalev, K. Aarma, A. Lehtmets, P. Ööpik, J. Lass, H. Hinrikus, Methods for classifying depression in single channel EEG using linear and nonlinear signal analysis, Computer Methods and Programs in Biomedicine 155 (2018) 11–17.

[42] E. Avots, K. Jermakovs, M. Bachmann, L. Päeske, C. Ozcinar, G. Anbarjafari, Ensemble approach for detection of depression using EEG features, Entropy 24 (2) (2022) 211.

[43] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, A. T. Campbell, Tracking depression dynamics in college students using mobile phone and wearable sensing, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2 (1) (2018) 1–26.

[44] Y. Rykov, T.-Q. Thach, I. Bojic, G. Christopoulos, J. Car, et al., Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling, JMIR mHealth and uHealth 9 (10) (2021) e24872.

[45] I. Moshe, Y. Terhorst, K. Opoku Asare, L. B. Sander, D. Ferreira, H. Baumeister, D. C. Mohr, L. Pulkki-Råback, Predicting symptoms of depression and anxiety using smartphone and wearable data, Frontiers in Psychiatry 12 (2021) 625247.

[46] K. Kipli, A. Kouzani, I. R. A Hamid, Investigating machine learning techniques for detection of depression using structural MRI volumetric

features, International Journal of Bioscience, Biochemistry and Bioinformatics 3 (5) (2013).

[47] M. Mousavian, J. Chen, Z. Traylor, S. Greening, Depression detection from sMRI and rs-fMRI images using machine learning, Journal of Intelligent Information Systems 57 (2021) 395–418.

[48] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Transactions on Knowledge and Data Engineering 32 (3) (2018) 588–601.

[49] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.

[50] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, H. Leung, Sensemood: depression detection on social media, in: Proceedings of ICMR'20 – the 2020 International Conference on Multimedia Retrieval, ACM, 2020, pp. 407–411.

[51] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, H. A. Schwartz, Facebook language predicts depression in medical records, Proceedings of the National Academy of Sciences 115 (44) (2018) 11203–11208.

[52] G. A. Bushnell, T. Stürmer, A. White, V. Pate, S. A. Swanson, D. Azrael, M. Miller, Predicting persistence to antidepressant treatment in administrative claims data: Considering the influence of refill delays and prior persistence on other medications, Journal of Affective Disorders 196 (2016) 138–147.

[53] M. Bertl, M. Shahin, P. Ross, D. Draheim, Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining, in: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 826–833. doi:10.1145/3555776.3577594. URL https://doi.org/10.1145/3555776.3577594

[54] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, Journal of Biomedical Informatics 83 (2018) 112–134.

[55] L. R. Medsker, L. Jain, Recurrent neural networks, Design and Applications 5 (2001) 64–67.

[56] K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[57] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (1) (2008) 61–80.

[58] J. Shang, T. Ma, C. Xiao, J. Sun, Pre-training of graph augmented transformers for medication recommendation, arXiv:1906.00346 [cs] (Nov. 2019).
URL http://arxiv.org/abs/1906.00346

[59] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[60] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342 (2019).

[61] A. Ashfaq, A. Sant'Anna, M. Lingman, S. Nowaczyk, Readmission prediction using deep learning on electronic health records, Journal of Biomedical Informatics 97 (2019) 103256.

[62] S. Hochreiter, J. Schmidhuber, et al., Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[63] C.-Y. Chi, S. Ao, A. Winkler, K.-C. Fu, J. Xu, Y.-L. Ho, C.-H. Huang, R. Soltani, Predicting the mortality and readmission of in-hospital cardiac arrest patients with electronic health records: a machine learning approach, Journal of Medical Internet Research 23 (9) (2021) e27798.

[64] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, A. Dai, Learning the graphical structure of electronic health records with graph convolutional transformer, in: Proceedings of AAAI'20 – the 34th AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 606–613.

[65] S. N. Shukla, B. M. Marlin, Interpolation-prediction networks for irregularly sampled time series, arXiv preprint arXiv:1909.07782 (2019).

[66] A. Wongkoblap, M. Vadillo, V. Curcin, Depression detection of Twitter posters using deep learning with anaphora resolution: Algorithm development and validation, JMIR Mental Health (2021).

[67] P. Mathur, R. Sawhney, S. Chopra, M. Leekha, R. Ratn Shah, Utilizing temporal psycholinguistic cues for suicidal intent estimation, in: Advances in Information Retrieval – Proceedings of ECIR'2020 : the 42nd European Conference on IR Research, Part II, Vol. 12036 of Lecture Notes in Computer Science, Springer, 2020, pp. 265–271.

[68] A. Nadeem, M. Naveed, M. Islam Satti, H. Afzal, T. Ahmad, K.-I. Kim, Depression detection based on hybrid deep learning sscl framework using self-attention mechanism: An application to social networking data, Sensors 22 (24) (2022) 9775.

[69] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, M. Uddin, Deep learning for depression detection from textual data, Electronics 11 (5) (2022) 676.

[70] G. Lam, H. Dongyan, W. Lin, Context-aware deep learning for multimodal depression detection, in: Proceeding of ICASSP'2019 – the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3946–3950.

[71] H.-S. Chiang, M.-Y. Chen, L.-S. Liao, Cognitive depression detection cyber-medical system based on EEG analysis and deep learning approaches, IEEE Journal of Biomedical and Health Informatics (2022).

[72] Y. Lin, B. N. Liyanage, Y. Sun, T. Lu, Z. Zhu, Y. Liao, Q. Wang, C. Shi, W. Yue, A deep learning-based model for detecting depression in senior population, Frontiers in Psychiatry 13 (2022).

[73] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of CVPR'2016 – the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[74] T. Guo, Z. Xu, X. Yao, H. Chen, K. Aberer, K. Funaya, Robust online time series prediction with recurrent neural networks, in: Proceedings of DSAA'2016 – the 2016 IEEE International Conference on Data Science and Advanced Analytic, IEEE, 2016, pp. 816–825.

[75] H. Jelodar, Y. Wang, R. Orji, S. Huang, Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach, IEEE Journal of Biomedical and Health Informatics 24 (10) (2020) 2733–2742.

[76] Y. Ho, S. Wookey, The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling, IEEE Access 8 (2019) 4806–4813.

[77] Estonian National Institute for Health Development, RA02: Residents with health insurance and health insurance coverage by sex and county – Tervisestatistika ja terviseuuringute andmebaas (2020).
URL `https://statistika.tai.ee/pxweb/en/Andmebaas/Andmebaas__04THressursid__12Ravikindlustatud/RA02.px`, `Lastaccessedon2022-03-04`

[78] X. Min, B. Yu, F. Wang, Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD, Scientific Reports 9 (1) (2019) 2362.

[79] S. Nolen-Hoeksema, Gender differences in depression, Current directions in psychological science 10 (5) (2001) 173–176.

[80] R. C. Kessler, H. Birnbaum, E. Bromet, I. Hwang, N. Sampson, V. Shahly, Age differences in major depression: results from the national comorbidity survey replication (ncs-r), Psychological medicine 40 (2) (2010) 225–237.

[81] R. T. Fielding, Representational State Transfer (REST), Chapter 5 in (R.T. Fielding): Architectural Styles and the Design of Network-based Software Architectures (Ph.D.). University of California, Irvine (2000).

[82] K. Kroenke, R. L. Spitzer, The PHQ-9: a new depression diagnostic and severity measure, Psychiatric Annals 32 (9) (2002) 509–515.

[83] S. A. Montgomery, M. Åsberg, A new depression scale designed to be sensitive to change, The British Journal of Psychiatry 134 (4) (1979) 382–389.

Appendix A. Number of Parameters per Model

Table A.6: Number of parameters per model

Models	# of Parameters
LR	1,649
CNN-LSTM	136,849
LSTM	101,871
FNN	5,156,331
GRU-decay	96,725
GRU-Δt	107,326
Att-GRU-decay	97,121

Appendix B. Data Distribution

Figure B.8: Age distribution

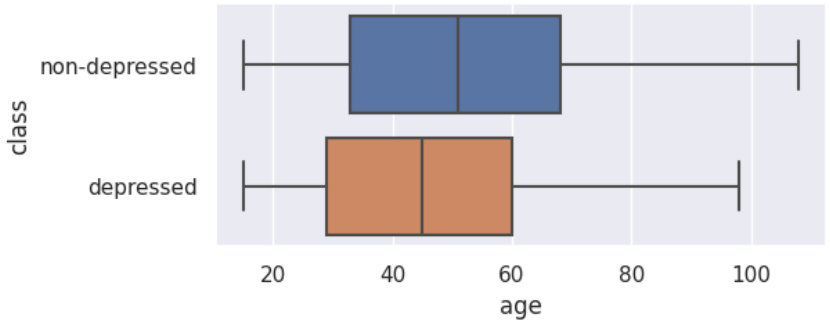


Figure B.9: Gender distribution

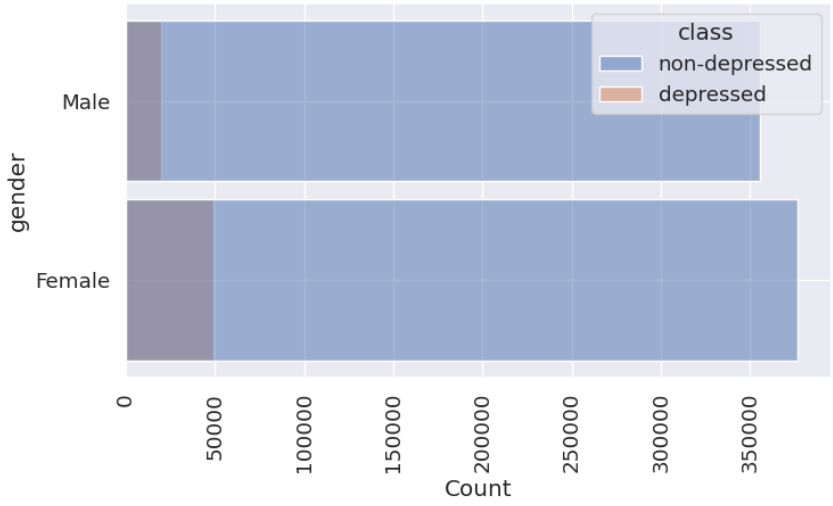


Figure B.10: Top 50 ICD-10 codes in our dataset

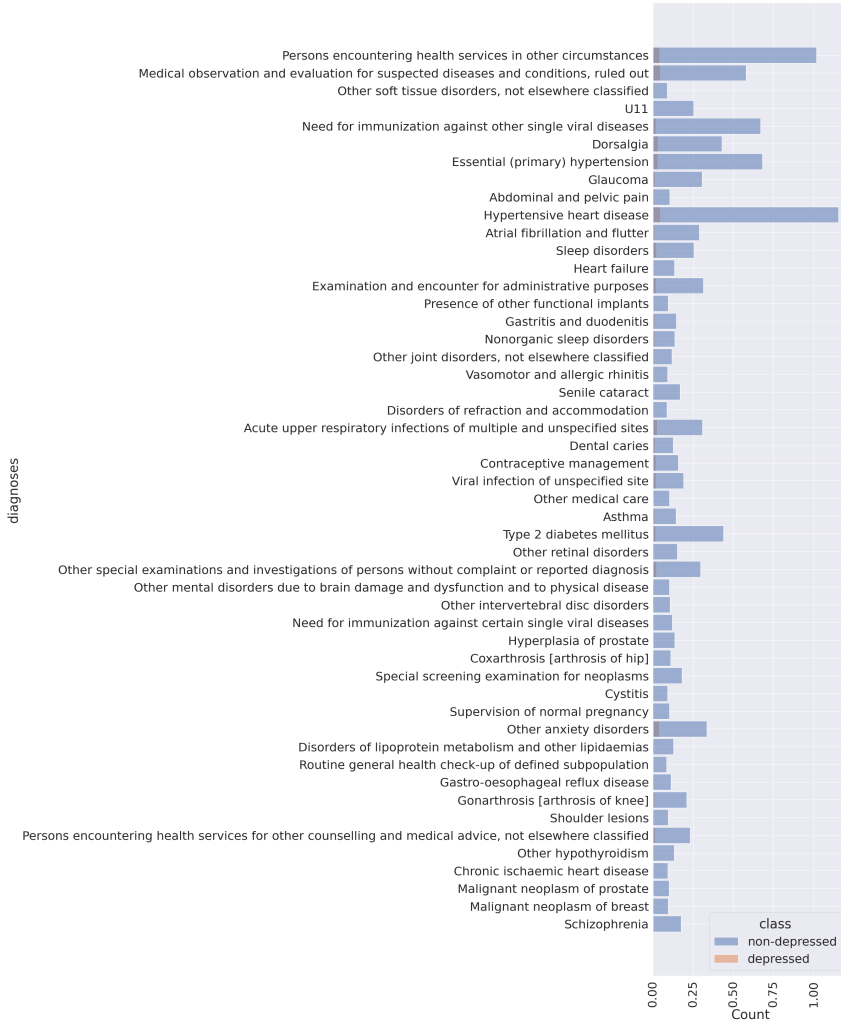


Figure B.11: Top 50 ICD-10 codes for patients with depression

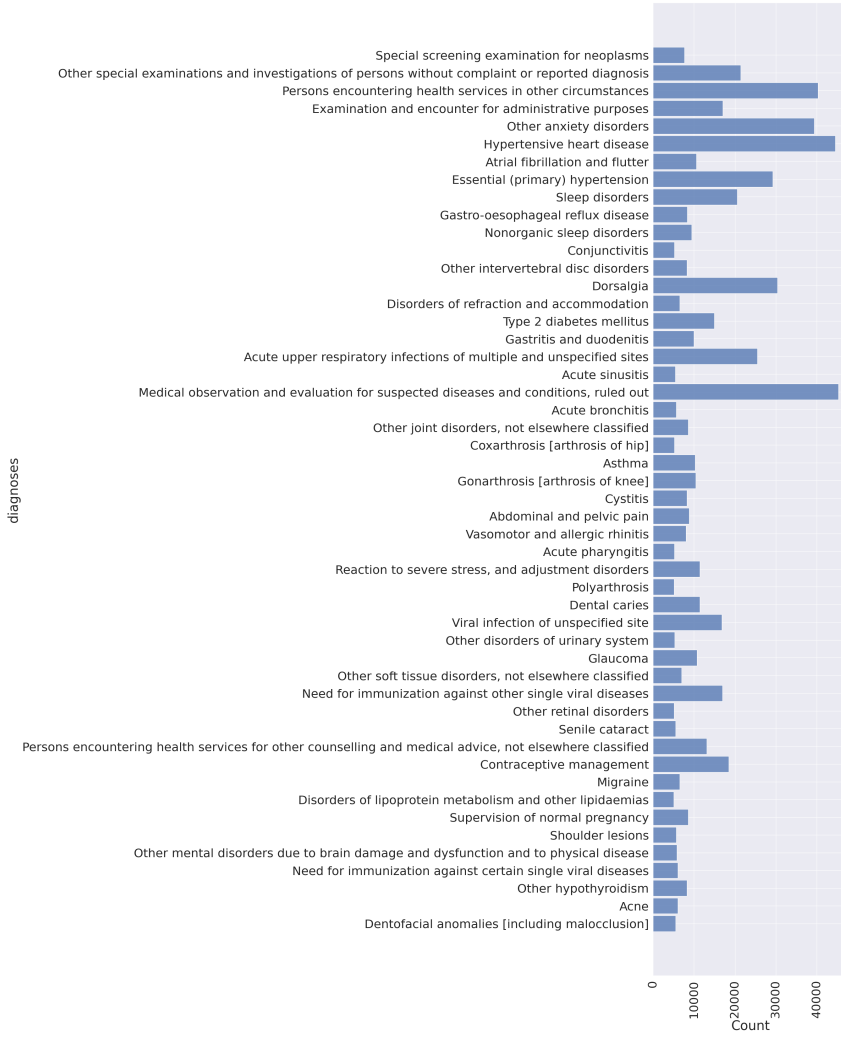
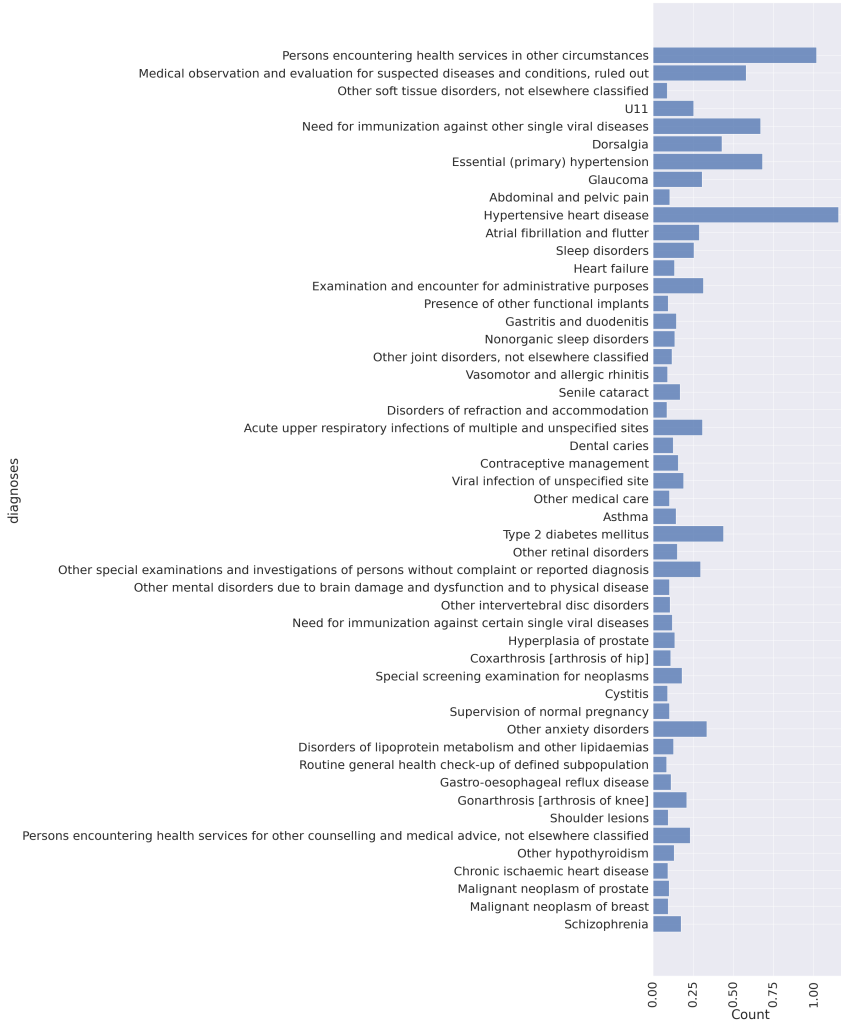


Figure B.12: Top 50 ICD-10 codes for patients without depression



Curriculum Vitae

1. Personal data

Name	Markus Bertl
Date and place of birth	29 April 1996, Vienna, Austria
Nationality	Austrain

2. Contact information

Address	Tallinn University of Technology, School of Information Technologies, Department of Health Technologies, Akadeemia tee 15a, 12618 Tallinn, Estonia
E-mail	mbertl@taltech.ee

3. Education

2019–...	Tallinn University of Technology, School of Information Technologies, Information and Communication Technology, PhD
2017–2019	St. Pölten University of Applied Sciences, Digital Healthcare, MSc.
2016–2017	University of Central Lancashire, Department of Computer Science, Computing, BSc. (hons)
2010–2015	Industrial College St. Pölten, Department of Informatics, Informatics, Ing.

4. Language competence

German	native
English	fluent
Estonian	basic level

5. Professional employment

2021–...	Unisys, IT Specialist
2016–2018	BIConcepts IT Consulting GmbH, Developer

6. Voluntary work

2021–...	Working group member to standardize AI ON-AG 001 42 - Artificial Intelligence (ISO/IEC JTC 001/SC 42, -SC 22, -SC 38)
2021–...	Working group member GAIA-X health. Building a next-gen data infrastructure as part of the European initiative GAIA-X
2021–...	Expert and Advisor in the European Civil Protection Mechanism
2015–...	Paramedic, International Disaster and Humanitarian Relief

7. Fields of research

- Artificial Intelligence
- Decision Support Systems
- Digital Health
- eGovernment

8. Scientific work

1. M. Bertl, J. Metsallik, and P. Ross. A Systematic Literature Review of AI-based Digital Decision Support Systems for post-traumatic Stress Disorder. *Frontiers in Psychiatry*, 13, 2022
2. M. Bertl, P. Ross, and D. Draheim. A Survey on AI and Decision Support Systems in Psychiatry – Uncovering a Dilemma. *Expert Systems with Applications*, 202:117464, 2022
3. M. Bertl, P. Ross, and D. Draheim. Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors. *Health Policy and Technology*, 2023
4. M. Bertl, K. J. I. Kankainen, G. Piho, D. Draheim, and P. Ross. Evaluation of Data Quality in the Estonia National Health Information System for Digital Decision Support. In *Proceedings of the 3rd International Health Data Workshop*. CEUR-WS, 2023
5. M. Bertl, P. Ross, and D. Draheim. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *Database and Expert Systems Applications*, pages 104–111. Springer International Publishing, 2021
6. M. Bertl, M. Shahin, P. Ross, and D. Draheim. Finding Indicator Diseases of Psychiatric Disorders in BigData Using Clustered Association Rule Mining. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 826–833. Association for Computing Machinery, 2023
7. M. Bertl, N. Bignoumba, P. Ross, S. B. Yahia, and D. Draheim. Evaluation of Deep Learning-based Depression Detection using Medical Claims Data. *SSRN*, 2023
8. K. Mak, H. C. Pilles, M. Bertl, and J. Klerx. Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie. *Schriftenreihe der Landesverteidigungsakademie. Wien: BM für Landesverteidigung und Sport*, 2018
9. M. Bertl. News Analysis for the Detection of Cyber Security Issues in Digital Healthcare: A Text Mining Approach to Uncover Actors, Attack Methods and Technologies for Cyber Defense. *Young Information Scientist*, 4:1–15, 2019
10. R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's Paradox: A Step Towards Fairness in Machine Learning. In *European Conference on Advances in Databases and Information Systems*, pages 67–76. Springer, 2022
11. M. Bertl, T. Klementi, G. Piho, P. Ross, and D. Draheim. How Domain Engineering Can Help to Raise Adoption Rates of Artificial Intelligence in Healthcare. In *Proceedings of the 25th International Conference on Information Integration and Web-based Applications & Services*. Springer Nature, 2023

Elulookirjeldus

1. Isikuandmed

Nimi	Markus Bertl
Sünniaeg ja -koht	29. Aprill 1996, Viin, Austria
Kodakondsus	Austria

2. Kontaktandmed

Aadress	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tervisetehnoloogiate instituut, Akadeemia tee 15a, 12618 Tallinn, Estonia
E-post	mbertl@taltech.ee

3. Haridus

2019–…	Doktorikraad, Infotehnoloogia teaduskond, Tallinna Tehnikaülikool
2017–2019	Magistrikraad, St. Pölten University of Applied Sciences, Digital Healthcare, MSc.
2016–2017	Bakalaureusekraad, University of Central Lancashire, Department of Computer Science, Computing, BSc. (hons)
2010–2015	Industrial College St. Pölten, Department of Informatics, Informatics, Ing.

4. Keelteoskus

saksa keel	emakeel
inglise keel	kõrgtase
eesti keel	põhitase

5. Teenistuskäik

2021–…	Unisys, IT Specialist
2016–2018	BIConcepts IT Consulting GmbH, Developer

6. Vabatahtlik töö

2021–…	AI standardimise töörühma liige ON-AG 001 42 - Artificial Intelligence (ISO/IEC JTC 001/SC 42, -SC 22, -SC 38)
2021–…	Töörühma liige GAIA-X health. Järgmise põlvkonna andmetaristu loomine Euroopa algatuse GAIA-X raames
2021–…	Euroopa kodanikukaitse mehhanismi ekspert ja nõunik
2015–…	Parameedik, rahvusvaheline katastroofi- ja humanitaarabi

7. Teadustöö põhisuunad

- Tehisintellekt
- Andmeanalüüs
- Digitaalsed otsustustoed
- Digitervis
- E-valitsus

8. Teadustegevus

Teadusartiklite, konverentsiteeside ja konverentsiettekannete loetelu on toodud ingliskeelse elulookirjelduse juures.

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-072-0 (PDF)