Natural language processing in healthcare: applications and value

Buchem, M.M. van

Natural Language Processing in Healthcare: Applications and Value

Marieke Meija van Buchem

Natural Language Processing in Healthcare: Applications and Value

Marieke Meija van Buchem

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Erwin Timmerman, persoonlijkproefschrift.nl
Cover: Reinier van Buchem

Natural Language Processing in Healthcare: Applications and Value

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 11 december 2024
klokke 10:00 uur
door

Marieke Meija van Buchem
geboren te Princeton, Verenigde Staten
in 1995

Promotor

Prof. Dr. E.W. Steyerberg

Copromotores

Dr. I.M.J. Kant	Universiteit Utrecht
Dr. M.P. Bauer	

Promotiecommissie

Prof. Dr. A.M. Stiggelbout	
Prof. Dr. M.R. Spruit	
Prof. Dr. S. Verberne	
Prof. Dr. A. Abu-Hanna	Universiteit van Amsterdam

‘It’s only complicated if you start thinking about it.”

Alan Watts

Table of contents

Chapter 1

Introduction

The healthcare sector is currently facing several critical challenges that threaten its accessibility and affordability. Among these challenges are personnel shortages, limited resources, and an ever-increasing demand for healthcare services. In response to these challenges, the healthcare sector has increasingly turned to technological innovations to find solutions. The most significant of these technological advancements has been the introduction of Electronic Health Records (EHRs). The shift from paper-based practices to EHRs in many hospitals aimed to streamline administrative processes and improve the efficiency of healthcare delivery. While the introduction of EHRs has transformed the landscape of healthcare data management, generating vast amounts of electronic, mostly unstructured healthcare data, it has not realized the anticipated reduction in administrative burdens[1]. This situation has paved the way for exploring further technological solutions to leverage the vast amount of data for enhancing healthcare services. Artificial Intelligence (AI), particularly through its subfield of Natural Language Processing (NLP), offers a promising avenue. In this thesis, applications and value of NLP are studied for healthcare.

1.1 Natural language processing

NLP combines linguistics and computer science, aiming to enable computers to process ('understand') natural language. NLP consists of many different subfields, with a wide variety of tasks and techniques. At the core of NLP are several key stages: data preprocessing, feature extraction, and modeling.

Data preprocessing involves several tasks aimed at enhancing the quality of the text and normalizing the text to improve the efficiency of subsequent processing steps. Common preprocessing techniques are shown in Figure 1. Which preprocessing techniques to apply depend on the NLP model that is used. For classical machine learning models such as logistic regression, support vector machines, and decision trees, more preprocessing steps are needed to reduce the dimensionality and sparsity of the data. More recent models are able to deal with the complexity of natural language and may only require tokenization.

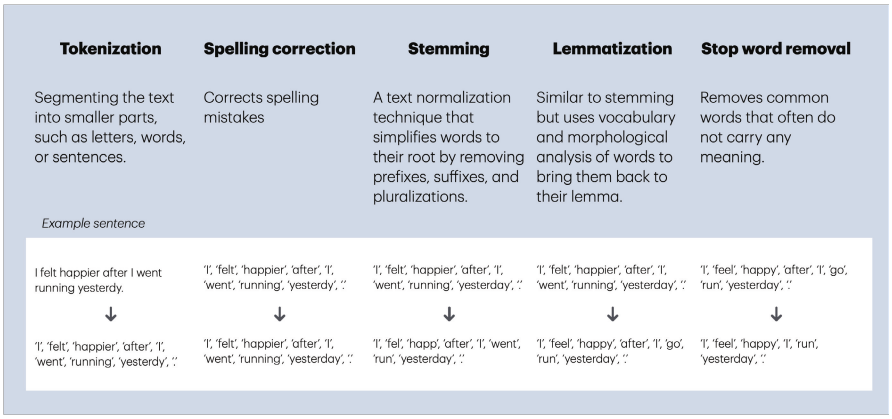


Figure 1: A visualization of common techniques for preprocessing in natural language processing[2].

Feature extraction is needed to transform the preprocessed data into a numerical format that can be analyzed by an algorithm. There are many different feature extraction techniques, with different levels of granularity (see Figure 2). Again, the model that is used establishes which feature extraction technique is suitable. As with preprocessing, this relates to the complexity the model can handle.

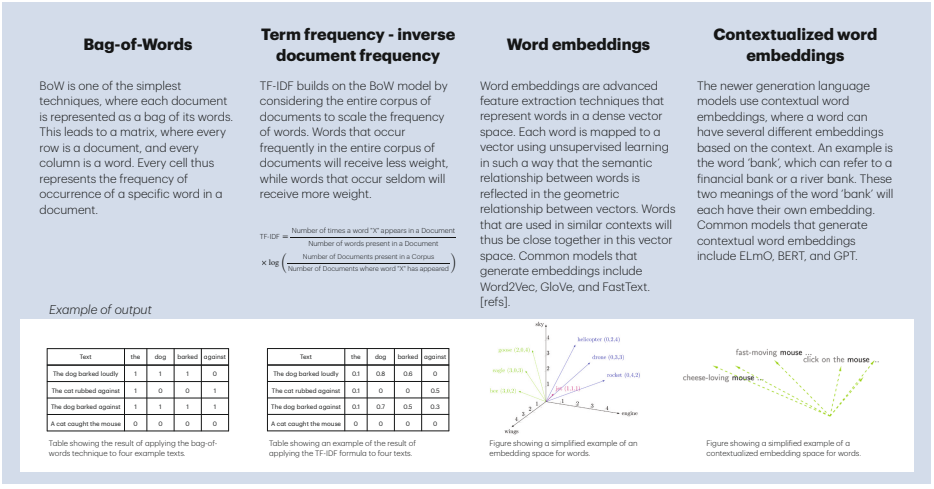


Figure 2: A visualization of common techniques for feature extraction in natural language processing, ordered by increasing complexity.

Modeling applies algorithms to the structured data that arises from the previous steps. There are many different modeling tasks and associated models, the ones presented here are those that emerged as most relevant for the settings discussed in the following chapters (see Figure 3).

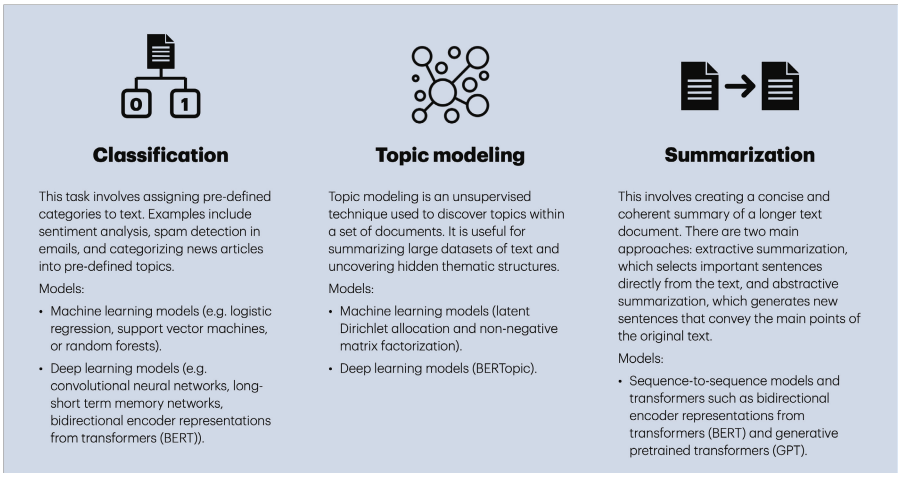


Figure 3: A visualization of the three natural language processing tasks included in this thesis.

Recent advances

In the past decade, large steps have been made in the advancement of the NLP field. The steady increase in computing power has made it possible to build language models with an increasing number of parameters. ELMo (Embeddings from Language Models), introduced in 2018, had approximately 90 million parameters[3]. Its successor, BERT (Bidirectional Encoder Representations from Transformers), introduced in 2018, has a base model with 110 million and a large model with 340 million parameters[4]. This was the first model to use a transformer architecture, a neural network architecture designed to be efficient at capturing relationships and dependencies between elements in a sequence, such as words in a sentence. This new architecture led to huge jumps in performance. The newest language models, such as OpenAI's GPT-3.5 (Generative Pretrained Transformers) and Google's PaLM (Pathways Language Model), also use the transformer architecture and have over 100 billion parameters, coining the new

term *large language models* (see Figure 4). With these new possibilities, a key question is how these techniques can be applied to healthcare data and if they can create value for clinical practice.

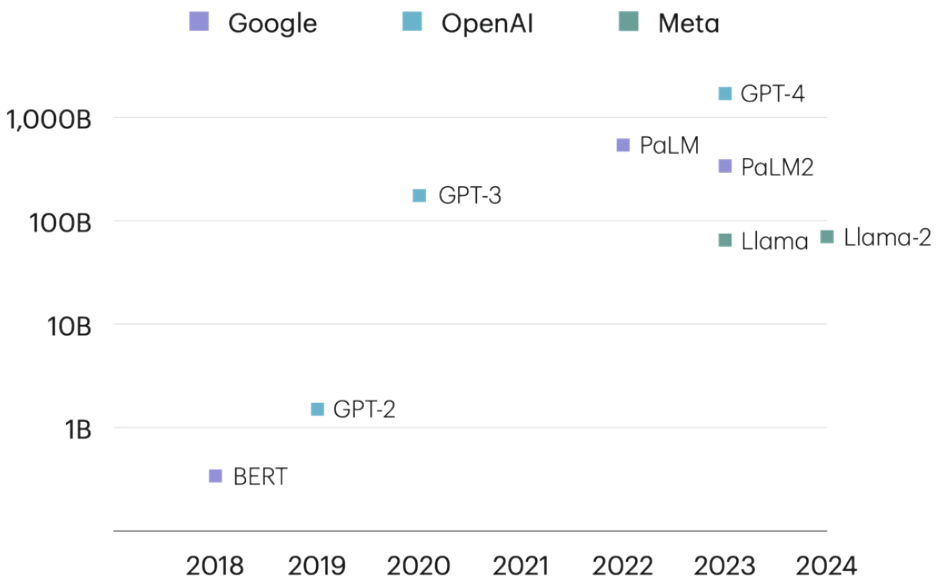


Figure 4: A timeline showing the (large) language models that have been introduced over the past five years and the number of parameters, plotted on a logarithmic scale.

1.2 Healthcare data

To create value at point-of-care, it is essential to use data routinely collected throughout the healthcare process. Routinely collected healthcare data for the purpose of NLP encompasses a broad spectrum of text-based information, including but not limited to:

- **Clinical data**, such as clinical notes, operative reports, and discharge summaries captured within the electronic health record (EHR). This data can be characterized as information from and to healthcare providers, summarizing interactions with patients, letters from and to other healthcare professionals, and notes from nurses, physicians, and other medical professionals.

- **Patient-generated data**, such as patient portal messages and patient experience surveys, characterized as information from patients to the hospital. With the increasing focus on patient-centered care, different efforts have been made to improve capturing the patient's voice and improving patient communication. This data is unique as it provides insights into patients' lives outside of the hospital.
- **Social media data**, from websites such as Reddit, Twitter, or online fora. The use of social media has increased over the past two decades, giving patients the opportunity to share their experiences, knowledge, and questions online. This data is characterized as peer-to-peer.

Previous studies show that free-text clinical data include valuable information not captured in structured data fields. Especially information about the complexity, evolving circumstances, uncertainty, and severity are often captured in free-text fields[5]. Furthermore, free-text clinical notes have previously been found to be more accurate, more reliable in identifying patients with certain diseases, and more understandable to review for other healthcare providers[6]. Similarly, for sources such as patient experience data, healthcare providers prefer answers to open-ended questions over closed-ended questions, as they provide more nuance[7]. All these examples clearly highlight the value of free-text healthcare data. With the volume of healthcare data estimated to grow with an annual rate of 36%[8], of which approximately 80% is unstructured, the need for the application of NLP is high.

1.3 Research questions

Despite the growing body of research on NLP in healthcare, the value of NLP for clinical practice is still unclear and implementation is lagging[9]. This thesis aims to clarify the value of the opportunities presented by NLP tools in clinical practice. It explores various applications of NLP within the healthcare setting and evaluates their value and offers insights into how NLP technologies can be effectively integrated into daily healthcare delivery to contribute to the improvement of healthcare accessibility and affordability. The following research questions will be addressed:

1. **Promising applications:** Which combination of NLP methods and data sources are most promising for enhancing healthcare delivery at the point of care?
2. **Challenges during development:** What are the principal challenges during the development of NLP models that hinder valuable adoption in clinical practice?
3. **Value for clinical practice**: What is the value of NLP applications for daily clinical practice?

1.4 Outline

The research questions will be addressed in the following chapters, which are divided into two parts. Part 1 focuses on the application of NLP in five different healthcare settings. Chapter 2 presents a scoping review about digital scribes in clinical practice. Chapter 3 through 6 describe the development of NLP models in different settings. Part 2 centers around the added value of NLP in clinical practice. Chapter 7 and 8 present pilot studies of applications presented in Part 1. See Table 1 for an overview of the clinical domains, data types and data sources. Figure 5 visualizes the outline and the relations between the chapters.

Table 1: overview of the chapters and their clinical domain, data type, and data source.

Chapter	Clinical domain	Data type	Data source
Part 1: Application of NLP in various healthcare settings			
2	General	Clinical	Scoping review
3	Oncology	Clinical	Electronic health record (Stanford)
4	Otorhinolaryngology	Patient-generated	Patient experience questionnaires (LUMC)
5	Oncology	Patient-generated	Patient portal messages, electronic health record (Stanford)
6	Cardiology	Social media	Online forum (Reddit)
Part 2: Evaluating the added value in clinical practice			
7	Otorhinolaryngology	Patient-generated	Patient experience questionnaires (LUMC)
8	Internal medicine	Clinical	Recorded mock conversations

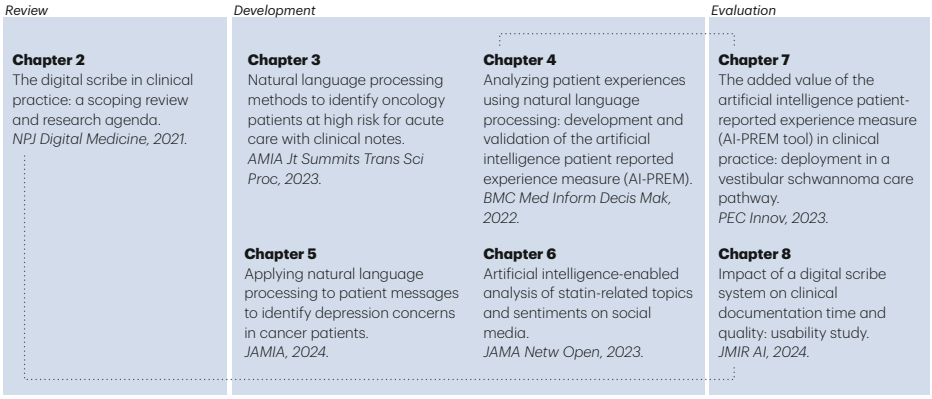


Figure 5: visualization of the outline of this thesis. Chapters that address the same setting are connected by a dotted line.

1.5 Terminology

Although NLP and the broader field of machine learning share many methodologies with biostatistics and epidemiology, different terms for similar concepts are used. In Table 2, key terminology used in this thesis is defined.

Table 2: definition of key terminology.

Term	Meaning
Model	A set of parameters and structure needed for a system to turn input data into output. Examples: logistic regression, support vector machines, neural network.
Label	The assigned category or value that an algorithm aims to predict, often used in supervised learning for training purposes.
Features	The individual measurable properties or characteristics of the data used by a model to make predictions. Similar to risk factors in epidemiology and independent variables in statistics.
Model training	Learning the association between features and labels. Similar to model fitting in statistics.

References

1. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W-J, Sinsky CA, Gilchrist VJ. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Medicine 2017;15(5):419–426.
2. Vajjala S, Majumder B, Gupta A, Surana H. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media, Inc.; 2020.
3. Peters, M. E. et al. Deep contextualized word representations. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 1, 2227–2237 (Association for Computational Linguistics, 2018).
4. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 1, 4171–4186 (Association for Computational Linguistics, 2019).
5. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Méd Ethics 2020;46(6):367–377.
6. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Méd Inform Assoc 2011;18(2):181–186.
7. Riiskjaer E, Ammentorp J, Kofoed P-E. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. Int J Qual Health C 2012;24(5):509–516.
8. Reinsel D, Gantz J, Rydning J. The Digitization of the World: From Edge to Core. IDC; 2018.
9. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(1):30–36.

Part 1

Application of Natural Language Processing in Various Healthcare Settings

Chapter 2

The Digital Scribe in Clinical Practice: A Scoping Review and Research Agenda

Marieke M. van Buchem, Hileen Boosman, Martijn P. Bauer, Ilse M.J. Kant, Simone A. Cammel & Ewout W. Steyerberg

npj Digital Medicine, 26 March 2021

2.1 Abstract

The number of clinician burnouts is increasing and has been linked to a high administrative burden. Automatic speech recognition (ASR) and natural language processing (NLP) techniques may address this issue by creating the possibility of automating clinical documentation with a "digital scribe". We reviewed the current status of the digital scribe in development towards clinical practice and present a scope for future research. We performed a literature search of four scientific databases (Medline, Web of Science, ACL Anthology, and Arxiv) and requested several companies that offer digital scribes to provide performance data. We included articles that describe the use of models on clinical conversational data, either automatically or manually transcribed, to automate clinical documentation. Of 20 included articles, three described ASR models for clinical conversations. The other 17 articles presented models for entity extraction, classification, or summarization of clinical conversations. Two studies examined the system's clinical validity and usability, while the other 18 studies only assessed their model's technical validity on the specific NLP task. One company provided performance data. The most promising models use context-sensitive word embeddings in combination with attention-based neural networks. However, the studies on digital scribes only focus on technical validity, while companies offering digital scribes do not publish information on any of the research phases. Future research should focus on more extensive reporting, iteratively studying technical validity and clinical validity and usability, and investigating the clinical utility of digital scribes.

2.2 Introduction

In the past few years, clinician burnout has become an acknowledged problem in healthcare. In a 2017 survey among 5000 US clinicians, 44% reported at least one symptom of burnout[1]. To investigate this problem, the National Academy of Medicine formed a committee focused on improving patient care by supporting clinician well-being. The committee's extensive report, called Taking Action Against Clinician Burnout, describes reasons for clinician burnout. An important reason is the increasing administrative burden[2]. Since the introduction of the electronic health record (EHR), the time spent on administrative tasks has increased to approximately half of a clinician's workday[3,4,5]. These administrative tasks decrease clinicians' career satisfaction[6] and negatively affect the clinician–patient relationship[7]. Other studies have assessed the relationship between EHR-use and burnout and found that more time spent on the EHR, especially after-hours, was linked to a higher risk of burnout[8,9].

Recently, clinicians have hired medical scribes to reduce the administrative burden, i.e., persons who manage administrative tasks, such as summarizing a consultation. Studies show positive results for the use of medical scribes, with clinicians spending more face-to-face time with patients and less after-hour time on the EHR[10,11]. Although a medical scribe might seem like the perfect solution, it shifts the burden to other personnel. As a result, direct medical costs increase, while the administrative burden remains substantial. Two recent perspectives[12,13] describe the need for a so-called digital scribe. This digital scribe uses techniques such as automatic speech recognition (ASR) and natural language processing (NLP) to automate (parts of) clinical documentation. The proposed structure for a digital scribe includes a microphone that records a conversation, an ASR system that transcribes this conversation, and a set of NLP models to extract or summarize relevant information and present it to the physician. The extracted information could, for instance, be used to create clinical notes, add billing codes, or use the extracted information for diagnosis support.

Companies like Google, Nuance, Amazon, and many startups are creating a digital scribe[14,15,16]. Although much needed, there are several concerns about

implementing a digital scribe in healthcare. These relate to technical aspects such as the accuracy of current ASR systems for transcription of spontaneous speech[13] and a digital scribe's ability to extract all the essential information from a non-linear, fragmented conversation[13,17]. There are also concerns related to a digital scribe's clinical utility, such as the effect on a physician's workflow. Such concerns need to be addressed before digital scribes can be safely implemented in practice. More specifically, successful implementation of an artificial intelligence (AI) tool, such as a digital scribe, requires a thorough investigation of its suitability, technical validity, clinical validity and usability, and clinical utility (see Box 1). A scoping review of current evidence is needed to determine the current status of the digital scribe and to make recommendations for future research.

Box 1: Four research phases

> Suitability: The first step aims to create a clear overview of the problem and find a suitable solution. In the digital scribe field, the problem is the administrative burden. Deciding on a suitable solution (e.g., symptom list, summary) is the next step towards determining the required model's output and a reliable ground truth[52]. When the problem and solution are clear, researchers can find a suitable dataset or collect data themselves. Researchers should also check if the dataset contains any unintended bias or underrepresented groups.
>
> Technical validity: Next, various methods may be created and the best performing model determined[55]. Apart from determining the model's overall performance, researchers should assess in which situations the model performs well and in which it performs less adequately. This includes assessing if the model performs consistently across different patient groups, for example gender[56]. The data source, model, and context in which the model was tested should all be described transparently[50]. Sharing data and code help the community better understand the models and enables researchers to build on past work[52].

Clinical validity and usability: Once the model passes the technical validation, the researchers should perform a qualitative evaluation of the output with the end-user. This evaluation has two goals: first, to evaluate whether the output makes sense and is clinically relevant; second, to evaluate how the output affects clinical practice. This includes the presentation of the output, the most appropriate timing, and the effect on end-users' decision making[57].

Clinical utility: In this last step, the researchers should prospectively study the model in clinical practice. First, the model might run in clinical practice without showing the output to the end-users. At specific time points, end-users analyze the output to identify any errors. If no new problems arise, a prospective study can be set up to determine clinical impact.

Objective

The purpose of the present study is to perform a scoping review of the literature and contact companies on the current status of digital scribes in healthcare. The specific research questions are:

- Which methods are being used to develop (part of) a digital scribe? (Suitability)
- How accurate are these methods? (Technical validity)
- Have any of these methods been evaluated in clinical practice? (Clinical validity and usability, clinical utility)

2.3 Methods

Data search

We performed a scoping review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) statement[18]. We searched Medline, Web of Science, Arxiv, and ACL Anthology for all relevant articles until December 25, 2020. Furthermore, we

scanned reference lists of relevant publications for additional articles. Search terms included terms describing the setting (clinical conversations) in combination with relevant methods (NLP, ASR) and usage of the output (clinical documentation). We also included "digital scribe" and "automated scribe" as search terms because these incorporate the setting, method, and goal. The full search queries can be found in Supplementary Table 1.

Besides, we aimed to include real-world data on existing digital scribes to bridge the gap between research and practice. Quiroz et al.[13] provided a list of active companies in the digital scribe space: Robin Healthcare, DeepScribe, Saykara, Sopris Health, Amazon, Nuance. These companies were requested to provide unpublished performance data for their digital scribe.

Inclusion and exclusion criteria

Our definition of a digital scribe is any system that uses a clinical conversation as input, either as audio or text, and automatically extracts information that can be used to generate an encounter note. We included articles that describe the performance of either ASR or NLP on clinical conversational data. A clinical conversation was defined as a conversation—in real life, over the phone, or via chat—between at least one patient and one healthcare professional. Because ASR and NLP are different fields of expertize and will often be described in separate studies, we chose to include studies that only focused on part of a digital scribe. Studies that described NLP models that were not aimed at creating an encounter note but, for example, extracted information for research purposes, were excluded. Articles written in any language other than English were excluded. Because of the rapidly evolving research field and the time lag for publications, proceedings and preprints were included.

Study selection

Two reviewers (M.M.v.B. and S.A.C.) independently screened all articles on title and abstract, using the inclusion and exclusion criteria. The selected articles were assessed for eligibility by reading the full text.

Data extraction and synthesis

The first reviewer extracted information from the included articles and the unpublished data provided by companies. The second reviewer verified the extracted information. The following aspects were extracted and assessed:

1. Setting and research phase
2. ASR models and performance
3. NLP tasks, models, and performance

2.4 Results

Study selection

Our search resulted in 2348 articles. After screening the titles and abstracts of these articles, we assessed 144 full-text articles for eligibility. We included 20 articles[19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38] for our analysis (Figure 1 and Supplementary Table 2). Of these, ten were conference proceedings[19,20,21,23,27,28,32,38], seven were workshop proceedings[22,26,29,34,35,36,37], two were journal articles[24,25], and three were Arxiv preprints[30,31,33].

Of the six contacted companies, DeepScribe[39] was the only one to provide unpublished data on their digital scribe system's performance. We were unable to obtain performance data from other companies.

Setting and research phase

Although all 20 studies aimed to decrease the administrative burden of clinical documentation in some way, the specific approaches and the setting differed greatly among studies. Three studies focused on improving the ASR for clinical conversations as the first step towards accurately extracting information from them[19,21,36]. Eleven studies chose to manually transcribe the conversations and performed NLP tasks on the transcripts[20,22,24,25,27,30,31,32,34,35,40]. Five studies used input data representative of the input of an implemented digital scribe (ASR transcripts or chat dialogs)[26,28,33,37,38].

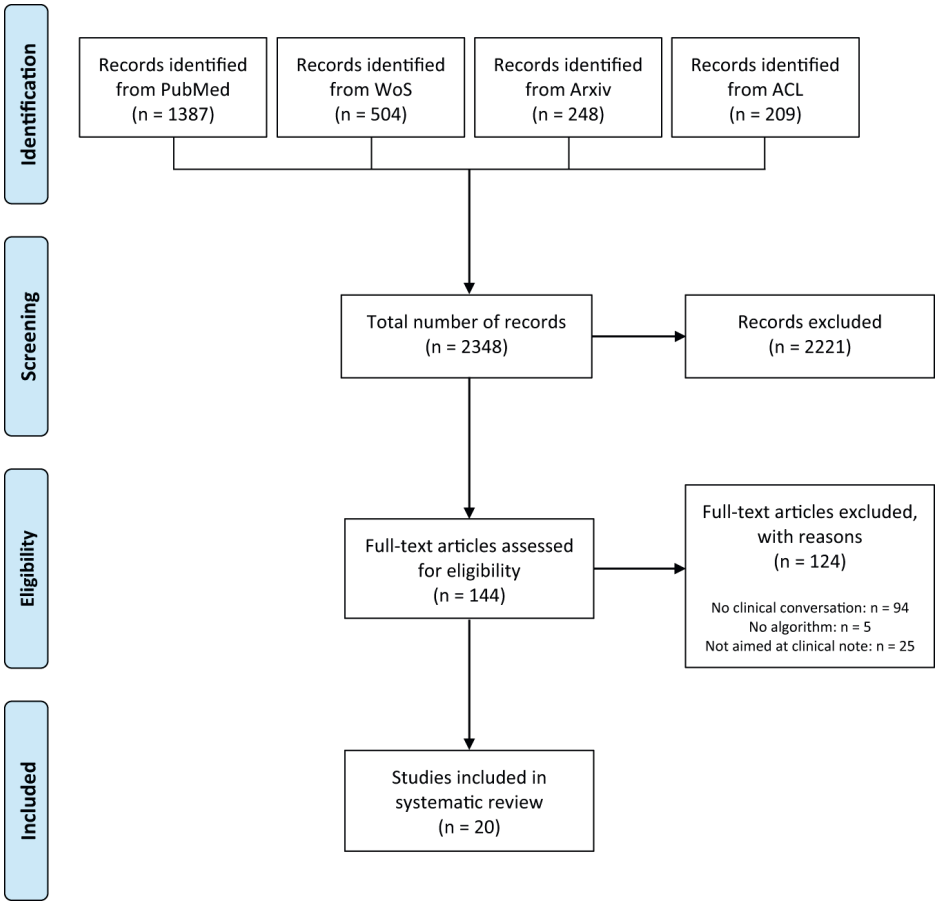


Figure 1: Inclusion flowchart. The four phases of article selection following the PRISMA-ScR statement.

Settings differed greatly between studies, as most did not define a specific specialty[19,21,22,23,26,31,32,33,34,35,36,38], while others were focused on primary care[20,25,27], home hemodialysis[24], orthopedic encounters[37], cardiology, family medicine, internal medicine[31], and patient-clinician dialogs via a telemedicine platform[28]. Fifteen studies were performed by or in collaboration with a company[19,20,21,23,25,26,27,28,30,33,34,35,36,37].

All included studies focused on the technical validity of the digital scribe; only two studies investigated the clinical validity and usability by performing a qual-

itative evaluation with end-users[24,28]. None of the studies investigated the clinical utility.

Automatic speech recognition (ASR)

In total, seven of 20 studies used ASR to automate clinical documentation[19,21,23,26,33,36,38], and one company provided data on their ASR system. Of these, two studies and the company presented a new ASR model[19,21], four used ASR to transcribe conversations as input for NLP models[26,33,37,38], one presented a model to correct ASR errors[36], and one compared the performance of existing ASR systems on clinical conversations[23] (see Supplementary Table 3).

In all studies, the metric used to evaluate the ASR transcripts was the word error rate (WER, see Box 2). The lowest WER was 14.1%, according to the unpublished data provided by DeepScribe. This ASR system combines Google Video Model[41], IBM Watson[42], and a custom-made Kaldi model[43]. The best performing published (as opposed to the unpublished data provided by DeepScribe) ASR system had a WER of 18%[19]. Four studies[23,26,33,36] used existing ASR systems and found WERs between 38% (IBM Watson) and 65% (Mozilla DeepSpeech[44]).

One study[36] presented a postprocessing model to correct ASR errors. By using an attention-based neural network, WERs were improved from 41 to 35% (Google Speech-to-Text[45]) and 36 to 35% (off-the-shelf open-source model[46]).

Box 2: Explanation of metrics

> *WER*: This metric counts the number of substitutions, deletions, and insertions in the automatic transcript, compared to the manual transcript. The lower the WER, the better the performance.
>
> *F1 score*: the F1 score is the harmonic mean between the precision (or positive predictive value) and the recall (or sensitivity).

ROUGE: this is a score that measures the similarity between the automatic summary and the golden standard summary, in unigrams (ROUGE-1), bigrams (ROUGE-2), or the longest common subsequence (ROUGE-L). The ROUGE-L score considers sentence-level structure, while the ROUGE-1 and ROUGE-2 scores only examine if a uni- or bigram occurs in both the automatic and golden standard summary.

Natural language processing (NLP) tasks and models

The NLP tasks that were performed could be split into three categories: entity extraction[20,25,26,27,30,32,35,38], classification[22,24,30,31,32,33,34,35], and summarization[22,24,28,29,31,37] (see Figure 2 and Supplementary Table 4). All except one study used word embeddings (see Box 3) as input to their model. This study did not use word embeddings as input but used a clustering model to create 2000 clusters[24]. The model's input consisted of the current words' clusters, the number of words, and the previous words' clusters.

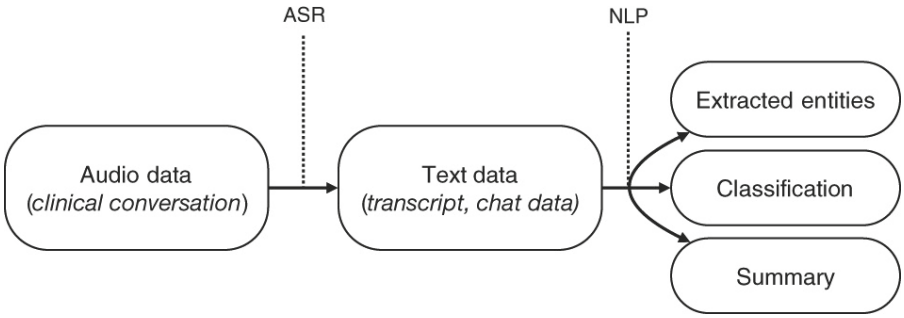


Figure 2: Overview of a digital scribe. Scope of the different aspects and techniques of the included digital scribes.

Entity extraction

The eight studies using entity extraction focused on extracting symptoms[20,25,27,32,38], medication regimen[20,26,27,32,35], and conditions[27]. However, the studies differed in the combination of entities and properties they extracted. Several studies examined the possibility of extracting symptoms and identifying whether a symptom was present or not[20,27,38],

while only one study focused on all the other combinations (i.e., medication dosage, frequency, symptom properties). Almost all studies reported their results as F1 scores (see Box 2). The tasks of extracting the medication, medication dosage, and symptom resulted in the highest F1 scores and thus showed the best performance (see Figure 3).

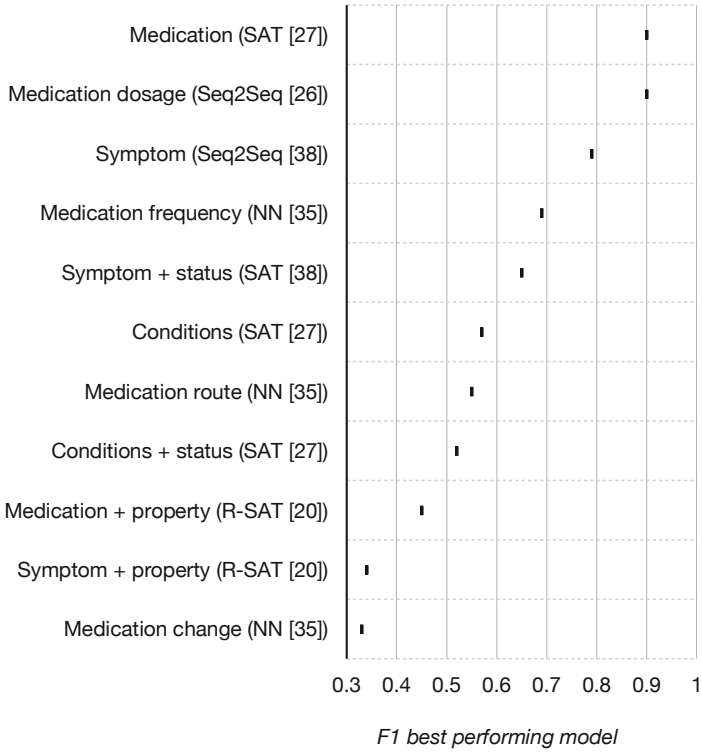


Figure 3: Performance of entity extraction models. Highest F1 scores per entity extraction task, with best performing model.

All studies used neural networks, although the type of neural network differed. Some studies used general neural networks[22,30,35], but most used neural network-based sequence models with attention (see Box 3). In the studies that compared different types of models, the neural networks with attention layer achieved higher F1 scores than the neural networks without attention layer (see Figure 3).

Three studies[27,32,38] performed an error analysis of which one investigated the symptoms that were incorrectly labeled as "absent". The authors reported that these symptoms were often discussed in multiple talk-turns. In the other study[27], ten human annotators categorized the cause of all labeling errors and the impact on the clinical note. They concluded that 16 to 32% of the errors did not affect the clinical note's content. Furthermore, most errors were caused by a failure of the model to take context into account or the lack of knowledge about a patient's medical background. In 29 to 42% of the errors, the human annotators agreed with the model, showing the difficulty of annotating the data. One study reported that most errors originated from informal language use and describing symptoms in physical manifestations ("I only sleep for 4 h")[32].

Two studies[26,38] made a comparison between manually transcribed and automatically transcribed data regarding the performance of their entity extraction model. Both found that models trained on manually transcribed data outperform the model trained on the automatically transcribed data. The difference in F1 for extracting symptoms was 0.79 versus 0.72, whereas the difference in ROUGE-1 (see Box 2) for extracting medication dosage was 85 versus 79[26].

Classification

Six studies performed a type of classification[22,24,30,33,34,35], which varied greatly: in which summary section it belonged[22,24,33,34]; if a sentence was said by the patient or the physician[33]; relevant diagnoses of the patient[22,30]; if any abnormalities were found in the medical history[30] (see Supplementary Table 4). A greater variety of models was used for classification than for entity extraction, although neural networks were used most often. The classification tasks resulting in the highest F1 scores were the classification of primary diagnosis, utterance type, and entity status (see Supplementary Table 4). In two of these tasks, support vector machines were used.

One study[33] tested their classification model on manually transcribed data and automatically transcribed data. The model performed better on the manually transcribed data, with a difference in F1 score ranging from 0.03 to 0.06, although they did not mention if the difference was significant.

One study assessed possible disparities of their classification model towards disadvantaged groups[34]. They formed 18 disadvantaged and advantaged groups based on gender, ethnicity, socioeconomic status, age, obesity, mental health, and location. In 7 of 90 cases, there was a statistically significant difference in favor of the advantaged group. The main reason for the disparity is a difference in the type of medical visit. For example, "blood" is a strong lexical cue to classify a sentence as important for the "Plan" section of the summary, but this word is said less often in conversations with Asian patients.

Summarization

Six studies[22,24,28,29,31,37] used NLP to summarize the conversation between patient and healthcare professional automatically. Four studies used pointer generator networks to create a hybrid extractive and abstractive summary[28,29,31,37]. One of these studies approached the summarization problem as a machine translation problem, where the transcript has to be "translated" to a summary[37]. This study compared the pointer generator network to three other attention-based models (see Supplementary Table 4).

The other two studies used extractive methods, where the output of the classification or entity extraction models was used to extract the most important utterances from the conversation[22,24]. The combination of these utterances formed the summary. One of these studies did not compare their summaries to a gold standard[24]; the other study asked physicians to extract the most important utterances as gold standard[22]. The F1 score for the latter study was 0.61.

All studies using pointer generator networks reported their results as ROUGE-scores. However, one study only reported their results as ROUGE-L relative error rate reduction[37], limiting the comparability with the other studies.

The ROUGE-L scores in the other three studies were 0.42[31], 0.55[28], and 0.55[29]. One study also presented a model that returned summaries with a ROUGE-L of 0.58, but this was based on manually extracting noteworthy utter-

ances[31]. When using the same model with automatically extracted noteworthy utterances, the performance dropped to 0.42.

The best performing model used a pretrained pointer generator network (see Box 3) fine-tuned on medical dialog summarization, with an added penalty for the generator distribution to force the model to favor copying text from the transcript over generating new text[28]. The other models were: a topic-aware pointer-generator network using embeddings (see Box 3)[29], which takes the topic of the current segment into account when copying or generating the next word; an LSTM architecture with BERT embeddings to extract noteworthy utterances (see Box 3)[31]; a combination of a transformer and pointer generator network that creates a summary per summary section (see Box 3)[37].

Two studies included physicians to evaluate their summaries[24,28]. One study examined physicians' ability to answer questions about patient care based on the automatic summary[24]. They did not find any significant difference in physicians' answers using the human-made summaries compared to the automatic summaries. Another study asked physicians to rate the amount of relevant information in the summaries[28]. Physicians found that 80% of the summaries included "all" or "most" relevant facts. The study did not specify which parts were deemed relevant or not or if the model missed specific information.

DeepScribe did not provide information on the models used for summarization but included how often a summary needed to be adjusted in practice. They report that 77% of their summaries do not need modification by a medical scribe before being sent to the physician. Furthermore, 74% of their summaries do not need modification from a medical scribe or a physician before being accepted as part of the patient's record, saving time on administrative tasks.

Box 3: Neural network-based sequence models with attention and word embeddings.

Attention-based neural networks: These models specifically take the sequence of the words into account, and have an attention layer. This layer acts as a filter, only passing the relevant subset of the input to the next layer.

Sequence2sequence (seq2seq)[52]: the seq2seq model uses a bidirectional encoder LSTM to include context, and has an attention mechanism to focus on the relevant parts of the input.

Span-attribute tagging model (SAT)[43]: the SAT model extracts symptoms and classifies them as present or not. It first identifies the relevant parts of the text and then classifies those relevant parts into symptoms that are or are not present. The relation-span-attribute tagging model (R-SAT) is a variant of the sat that focuses on relations between attributes.

Pointer generator network (PGNet)[53]: pgnets are based on the seq2seq architecture. The added value of a PGNet is that it has the ability to generate new words or copy words from the text, increasing the summary's accuracy.

Word embeddings: word embeddings are used to numerically represent words in a way that similar words have similar representations. For example, the words 'physician', 'clinician', and 'doctor' will have similar representations. There are different types of word embeddings, but the most important distinction for this review is between context-sensitive and context-insensitive embeddings. Context-sensitive embeddings have different representations for words that have multiple meanings. For example, the word 'bank' can mean a riverbank, or a financial institution. Some word embeddings, like Word2Vec[54], allow only one representation per word, whereas context-sensitive embeddings like ELMo[55] and BERT[56] can distinguish the different meanings of the word 'bank'.

2.5 Discussion

This scoping review provides an overview of the current state of the development, validation, and implementation of digital scribes. Although the digital scribe is still in an early research phase, there appears to be a substantial research body testing various techniques in different settings. The first results are promising: state-of-the-art models are trained on vast corpora of annotated clinical conversations. Although the performance of these models varies per task, the results give a clear view of which tasks and which models yield high performance. Reports of clinical validity and usability, and especially clinical utility are, however, mostly lacking.

All studies focusing on ASR used physician–patient dialogs without further specification of the setting. In general, existing ASR systems not explicitly trained on clinical conversations did not perform well, with WERs up to 65%. The speech recognition systems trained on thousands of clinical conversations had WERs as low as 18%. This WER is still high compared to the claimed WERs of general, state-of-the-art, available ASR systems that attain WERs as low as 5%[47]. The difference in performance can be explained by the uncontrolled setting of clinical conversations with background noise, multiple speakers, and the spontaneity of the speech[13]. However, these aspects were not reported by any of the studies, complicating the comparison of WERs. Two new approaches decreased the WER by postprocessing the automatic transcript[36] and combining multiple ASR systems (DeepScribe). These approaches are promising new ways to decrease the WER. However, what is most important is whether the WER is good enough to extract all the relevant information. Currently, the NLP models trained on manually transcribed data outperform those trained on automatically transcribed data, which means there is room for improvement of the WER.

When comparing the different NLP tasks, the diverseness in both tasks and underlying models was large. The classification models focused mainly on extracting metadata, such as relevance or structure induction of an utterance, and used various models ranging from logistic regression to neural networks. The entity extraction models were more homogeneous in models but extracted many

different entities, complicating the comparison, whereas the summarization task was mostly uniform, both in models and in metrics. One notable aspect of the NLP tasks overall is the use of word embeddings. Only one study did not use word embeddings, but this was a study from 2006 when context-sensitive word embeddings were not yet available. All the other studies were published after 2019 and used various word embeddings as input. The introduction of context-sensitive word embeddings has been essential for extracting entities and summarizing clinical conversations.

In the entity extraction task, the specific tasks, such as extracting symptoms, led to better performance than more general tasks, such as extracting symptoms and their properties. An explanation for this is the heterogeneity in, for example, symptom properties, which entail the location, severity, duration, and other characteristics of a symptom. These properties can be phrased in various ways, in contrast to medication or frequency, which will be much more homogeneous in phrasing. Therefore, this homogeneity leads to many more annotations per entity, increasing performance.

The same pattern was observed in the models, where the addition of an attention layer increased performance. This finding is in line with previous studies on neural attention[48,49], which describe the decrease in neural networks' performance with increased input length. By adding weights to the input text, the model knows which parts of the text are important for its task. Adding attention not only improves performance; it also decreases the amount of training data needed, which is useful in a field such as healthcare, where gathering large datasets can be challenging.

In the studies performing the entity extraction task, the error analyses showed that often, symptoms, medications, or properties are hard to interpret even by human annotators. This result is in line with the concerns discussed in the introduction, questioning if a model would accurately extract all relevant information from a non-linear, fragmented conversation. However, this takes the concern one step further, namely how the "gold standard" will be determined if there is ambiguity between human annotators. More research is needed to define

methods for developing gold standards. Shafran et al.[27] have taken an exciting first step towards such a method by publishing an article about the development of their corpus, including how they dealt with ambiguity and labeling errors.

The studies investigating summarization of the clinical conversation used both extractive and abstractive summarization techniques. However, the extractive techniques resulted in a list of the most important utterances instead of a new, full summary. Therefore, the studies performing abstractive summarization are more interesting to discuss. All four studies used the same model, the pointer generator network[28,29,31,37]. This network's advantage, especially with the studies' additions, makes sure it copies more words than it generates, keeping the summary as close to the conversation as possible. Two studies also included a quality check by physicians, which gives more insight into the possibility of implementation[24,28]. However, it would have been interesting to include error analyses to investigate the models' blind spots.

Future work

First of all, we believe it is vital to improve the ASR for clinical conversations further and use them as input for NLP models. A remarkable finding was that most studies used manually transcribed conversations as input to their NLP model. These manual transcripts may outperform automatically transcribed conversations regarding data quality, leading to an overestimation of the results. NLP models that require manual transcription may increase administrative burden when implemented in clinical practice.

Secondly, the current body of research is mostly focused on improving the performance of different models. Although some studies performed error analyses and qualitative analyses of the model's output, most did not. Moreover, most studies did not fully cover the technical validity phase because of insufficient reporting on the setting, data, and situations in which the model succeeded and failed. This information is essential to describe for a model that could potentially be implemented in clinical practice. The proposed models might contain bias or lead to unintended results, as Ferracane and Konam[34,37] show. This study is an inspirational example of how researchers can investigate the strengths and

weaknesses of their model. A recent paper by Hernandez-Boussard et al.[50] proposes reporting standards for AI in healthcare, which should be the basis for reporting on digital scribes as well.

Although most studies are in an early development phase, including qualitative analyses of the model's output is necessary to know if the solution researchers or developers are working on is applicable in practice. The lack of implementation following the development of an AI model is common in healthcare[51], which can be limited by investigating clinical validity and usability while working on technical validity. A good example is the study by Joshi et al.[28], where physicians qualitatively analyze the model's output. These results lead to new insights for improving technical validity. Studying these two research phases iteratively leads to a solution that is well-suited for clinical practice.

Most of the presented models need to be technically and clinically validated before moving on to the clinical utility phase. However, the companies already offering digital scribes seem to have skipped all four research phases, including clinical utility. We urge these companies to publish data on their digital scribes' technical validity, clinical validity and usability, and clinical utility. Not only is transparency in the model and its performance crucial for clinical practice, but it also helps the community better understand the models and enables researchers to build on past work[52].

The suitability phase falls outside the scope of this review but is nevertheless vital for developing and implementing the digital scribe. One research group has published several studies investigating which parts of a clinical conversation are relevant for creating a summary and how physicians see the potential role of a digital scribe[53,54]. These studies should be the starting point for researchers and developers working on a digital scribe.

Strengths and limitations

The current work is the first effort to review all available literature on developing a digital scribe. We believe our search strategy was complete, leading to a comprehensive and focused scope of the digital scribe's current research body.

By adding the company's data, we create a broader overview than just the digital scribe's scientific status. However, this data is unpublished, which means we have to trust the company in providing us with legitimate data. We hope this review is an encouragement for other companies to study their digital scribes scientifically.

One limitation is the small number of journal papers included in this review, as opposed to the amount of Arxiv preprints and workshop proceedings. These types of papers are often refereed very loosely. However, only including journal papers would not lead to a complete scope of this quickly evolving field.

Contacting various digital scribe companies was a first step towards gaining insight into implemented digital scribes and their performance on the different ASR and NLP tasks. Although only one company replied, we believe it is a valuable addition to this review. It indicates that their implemented digital scribe does not differ significantly in techniques or performance from the included studies' models while already saving physicians' time. Nevertheless, it highlights the gap between research and practice. The studies published by companies all describe techniques that are not part of a fully functional digital scribe (yet). However, none of the companies offering digital scribes have published about the technical validity, clinical validity and usability, or clinical utility of their systems.

Conclusion

Although the digital scribe field has only recently started to accelerate, the presented techniques achieve promising results. The most promising models use context-sensitive word embeddings in combination with attention-based neural networks. However, the studies on digital scribes only focus on technical validity, while companies offering digital scribes do not publish on any of the research phases. Future research should focus on more extensive reporting, iteratively studying technical validity and clinical validity and usability, and investigating the clinical utility of digital scribes.

Data availability

Any data generated or analyzed are included in this article and the Supplementary Information files. Aggregate data analyzed in this study are available from the corresponding author on reasonable request.

Ethics declarations

The authors declare no competing interests.

Supplementary Information

References

1. Shanafelt, T. D. et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. Mayo Clin. Proc. 94, 1681–1694 (2019).
2. National Academies of Sciences, Engineering, and Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being (The National Academies Press, 2019).
3. Arndt, B. G. et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann. Fam. Med. 15, 419–426 (2017).
4. Sinsky, C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann. Intern. Med. 165, 753–760 (2016).
5. Tai-Seale, M. et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff. 36, 655–662 (2017).
6. Rao, S. K. et al. The impact of administrative burden on academic physicians. Acad. Med.92, 237–243 (2017).
7. Pelland, K. D., Baier, R. R. & Gardner, R. L. "It's like texting at the dinner table": a qualitative analysis of the impact of electronic health records on patient-physician interaction in hospitals. J. Innov. Health Inform. 24, 216–223 (2017).
8. Robertson, S. L., Robinson, M. D. & Reid, A. Electronic health record effects on work-life balance and burnout within the I3 population collaborative. J. Grad. Med. Educ. 9, 479–484 (2017).
9. Gardner, R. L. et al. Physician stress and burnout: the impact of health information technology. J. Am. Med. Inform. Assoc. 26, 106–114 (2019).
10. Gidwani, R. et al. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial. Ann. Fam. Med. 15, 427–433 (2017).
11. Mishra, P., Kiang, J. C. & Grant, R. W. Association of medical scribes in primary care with physician workflow and patient experience. JAMA Intern. Med. 178, 1467 (2018).
12. Coiera, E., Kocaballi, B., Halamka, J. & Laranjo, L. The digital scribe. Npj Digital Med. 1, 1–5 (2018).
13. Quiroz, J. C. et al. Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digital Med. 2, 1–6 (2019).
14. Ambient clinical intelligence: the exam of the future has arrived. Nuance Communications (2019). Available at: https://www.nuance.com/healthcare/ambient-clinical-intelligence.html. (Accessed: 18th February 2021).
15. Amazon comprehend medical. Amazon Web Services, Inc (2018). Available at: https://aws.amazon.com/comprehend/medical/. (Accessed: 18th February 2021).
16. Robin Healthcare | automated clinic notes, coding and more. Robin Healthcare (2019). Available at: https://www.robinhealthcare.com. (Accessed: 18th February 2021).
17. Lin, S. Y., Shanafelt, T. D. & Asch, S. M. Reimagining clinical documentation with artificial intelligence. Mayo Clin. Proc. 93, 563–565 (2018).

18. Tricco, A. C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann. Intern. Med. 169, 467–473 (2018).

19. Chiu, C.-C. et al. Speech recognition for medical conversations. Proc. Interspeech 2018, 2972–2976 (2018).

20. Du, N., Wang, M., Tran, L., Li, G. & Shafran, I. Learning to infer entities, properties and their relations from clinical conversations. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 4979–4990 (Association for Computational Linguistics, 2019).

21. Shafey, L. E., Soltau, H. & Shafran, I. Joint speech recognition and speaker diarization via sequence transduction. Proc. Interspeech 2019, 396–400 (2019).

22. Jeblee, S., Khattak, F. K., Crampton, N., Mamdani, M. & Rudzicz, F. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In Proc. of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI), 65–74 (Association for Computational Linguistics, 2019).

23. Kodish-Wachs, J., Agassi, E., Kenny, P. & Overhage, J. M. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In Proc. of the Annual AMIA Symposium, 683–689 (American Medical Informatics Association, 2018).

24. Lacson, R. C., Barzilay, R. & Long, W. J. Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. J. Biomed. Inform.39, 541–555 (2006).

25. Rajkomar, A. et al. Automatically charting symptoms from patient-physician conversations using machine learning. JAMA Intern. Med. 179, 836 (2019).

26. Selvaraj, S. P. & Konam, S. Medication regimen extraction from medical conversations. In Proc. of International Workshop on Health Intelligence of the 34th AAAI Conference on Artificial Intelligence (Association for Computational Linguistics, 2020).

27. Shafran, I. et al. The medical scribe: corpus development and model performance analyses. In Proc. of the 12th Language Resources and Evaluation Conference (European Language Resources Association, 2020).

28. Joshi, A., Katariya, N., Amatriain, X. & Kannan, A. Dr. summarize: global summarization of medical dialogue by exploiting local structures. In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3755–3763 (Association for Computational Linguistics, 2020).

29. Liu, Z., Ng, A., Lee, S., Aw, A. T. & Chen, N. F. Topic-aware pointer-generator networks for summarizing spoken conversations. In Proc. IEEE Automatic Speech Recognition Understanding Workshop 2019, 814–821 (IEEE, 2019).

30. Krishna, K., Pavel, A., Schloss, B., Bigham, J. P. & Lipton, Z. C. Extracting Structured Data from Physician-Patient Conversations by Predicting Noteworthy Utterances. in Shaban-Nejad A., Michalowski M., Buckeridge D.L. (eds) Explainable AI in Healthcare and Medicine. Studies in Computational Intelligence, vol 914 (Springer International Publishing, 2021).

31. Krishna, K., Khosla, S., Bigham, J. P. & Lipton, Z. C. Generating SOAP notes from doctor-patient conversations. Preprint at arXiv (2020).

32. Khosla, S., Vashishth, S., Lehman, J. F. & Rose, C. MedFilter: improving extraction of task-relevant utterances through integration of discourse structure and ontological knowledge. In Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7781–7797 (Association for Computational Linguistics, 2020).

33. Schloss, B. & Konam, S. Towards an automated SOAP note: classifying utterances from medical conversations. Preprint at arXiv (2020).

34. Ferracane, E. & Konam, S. Towards fairness in classifying medical conversations into SOAP sections. In To be presented at AAAI 2021 Workshop: Trustworthy AI for Healthcare (AAAI Press, 2020).

35. Patel, D., Konam, S. & Selvaraj, S. P. Weakly supervised medication regimen extraction from medical conversations. In Proc. of the 3rd Clinical Natural Language Processing Workshop, 178–193 (Association for Computational Linguistics, 2020).

36. Mani, A., Palaskar, S. & Konam, S. Towards understanding ASR error correction for medical conversations. In Proc. of the First Workshop on Natural Language Processing for Medical Conversations, 7–11 (Association for Computational Linguistics, 2020).

37. Enarvi, S. et al. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In Proc. of the First Workshop on Natural Language Processing for Medical Conversations, 22–30 (Association for Computational Linguistics, 2020).

38. Du, N. et al. Extracting symptoms and their status from clinical conversations. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 915–925 (Association for Computational Linguistics, 2019).

39. DeepScribe - AI-Powered Medical Scribe. DeepScribe (2020). Available at: https://www.deepscribe.ai. (Accessed 18th February 2021).

40. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat. Med. 26, 1364–1374 (2020).

41. Transcribing videos | Cloud speech-to-text documentation. Google Cloud (2016). Available at: https://cloud.google.com/speech-to-text/docs/video-model. (Accessed 18th February 2021).

42. Watson speech to text - Overview. IBM (2021). Available at: https://www.ibm.com/cloud/watson-speech-to-text. (Accessed 18th February 2021).

43. Kaldi ASR. Kaldi (2015). Available at: https://kaldi-asr.org. (Accessed 18th February 2021).

44. mozilla/DeepSpeech. GitHub (2020). Available at: https://github.com/mozilla/DeepSpeech. (Accessed 18th February 2021).

45. Speech-to-text: automatic speech recognition | Google Cloud. Google Cloud (2016). Available at: https://cloud.google.com/speech-to-text. (Accessed 18th February 2021).

46. Peddinti, V. et al. Jhu aspire system: robust LVCSR with TDNNs, Ivector adaptation and RNN-LMs. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 539–546 (IEEE, 2015).

47. Hu, K., Sainath, T. N., Pang, R. & Prabhavalkar, R. Deliberation model based two-pass end-to-end speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7799–7803 (IEEE 2020).

48. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. Preprint at arXiv (2014).

49. Cho, K. et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734 (Association for Computational Linguistics, 2014).
50. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A. & Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J. Am. Med. Inform. Assoc. 27, 2011–2015 (2020).
51. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. Nat. Med. 25, 30–36 (2019).
52. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340 (2019).
53. Kocaballi, A. B. et al. Envisioning an artificial intelligence documentation assistant for future primary care consultations: a co-design study with general practitioners. J. Am. Med. Inform. Assoc. 27, 1695–1704 (2020).
54. Quiroz, J. C. et al. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. Health Inform. J. 26, 2906–2914 (2020).
55. Larson, D. B. et al. Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging models: summary and recommendations. J. Am. Coll. Radiol. 18, 413–424 (2020).
56. Tatman, R. Gender and dialect bias in YouTube's automatic captions. In Proc. of the First ACL Workshop on Ethics in Natural Language Processing, 53–59 (Association for Computational Linguistics, 2017).
57. Vasey, B. et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat. Med. 27, 186–187 (2021).
58. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In Proc. of the 27th International Conference on Neural Information Processing Systems (NIPS) 2, 3104–3112 (MIT Press, 2014).
59. See, A., Liu, P. J. & Manning, C. D. Get to the point: summarization with pointer-generator networks. In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, 1, 1073–1083 (Association for Computational Linguistics, 2017).
60. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In Proc. of the 26th International Conference on Neural Information Processing Systems (NIPS) 2, 3111–3119 (Curran Associates Inc., 2013).
61. Peters, M. E. et al. Deep contextualized word representations. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 1, 2227–2237 (Association for Computational Linguistics, 2018).
62. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 1, 4171–4186 (Association for Computational Linguistics, 2019).

Chapter 3

Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes

Claudio Fanconi, Marieke M. van Buchem,
Tina Hernandez-Boussard

Proceedings – AMIA Joint Summits on Translational Science,
June 16 2023

3.1. Abstract

Clinical notes are an essential component of a health record. This paper evaluates how natural language processing (NLP) can be used to identify the risk of acute care use (ACU) in oncology patients, once chemotherapy starts. Risk prediction using structured health data (SHD) is now standard, but predictions using free-text formats are complex. This paper explores the use of free-text notes for the prediction of ACU in lieu of SHD. Deep Learning models were compared to manually engineered language features. Results show that SHD models minimally outperform NLP models; an l1-penalised logistic regression with SHD achieved a C-statistic of 0.748 (95%-CI: 0.735, 0.762), while the same model with language features achieved 0.730 (95%-CI: 0.717, 0.745) and a transformer-based model achieved 0.702 (95%-CI: 0.688, 0.717). This paper shows how language models can be used in clinical applications and underlines how risk bias is different for diverse patient groups, even using only free-text data.

3.2 Introduction

Oncology patients undergoing chemotherapy often need acute care utilisation (ACU) and hospitalisation after starting chemotherapy. These interventions account for nearly half of the costs associated with oncology care in the United States[1,2]. Evidence suggests roughly 50% of these healthcare encounters are potentially preventable through early outpatient interventions[3,4]. A previous paper by Peterson et al[5] introduced a machine learning (ML) model, using structured health data (SHD) from electronic health records (EHR), to identify patients at high risk for ACU after chemotherapy initiation. In total, they trained their model using 760 inputs and retained 125 to predict the risk of ACU. This work, and others, highlight the potential of data-driven models to predict ACU risk using SHD[6–8].

However, most EHRs are not mapped to a common data model and they are not necessarily standardised between different facilities. To replicate other hospital's predictive models they could require intensive data preparation. On the other hand, 96% of hospitals in the US collect digital clinical notes from physicians and nurses in 2019[9]. Natural language processing (NLP) methods can extract useful information from these unstructured clinical texts.

NLP methods have already proven useful in clinical applications, e.g. for predicting critical care outcomes in intensive care units[10], classifying procedures and diagnoses[11], or predicting outcomes after an ischemic stroke12. In particular, deep learning-based language models have become popular in recent years[13], being used to identify 30-day hospital readmissions[14,15] or Statin non-use[16].

This study aims to assess the added value for identifying patients at risk of needing ACU by replacing tabular inputs with features from unstructured clinical notes or by combining both modalities. Second, we aim to investigate whether novel deep learning language models outperform traditional language feature extraction and linear models. We investigated these aspects by developing five

predictive models of ACU risk trained with different inputs and compared their predictive performance and utility when used at the point of care.

3.3 Methods

Data Collection

In 2019, the Centers for Medicare & Medicaid Services (CMS) introduced the Chemotherapy Measure (also referred to as OP-35), a quality measure that captures hospital admissions or emergency department visits of adult patients related to potentially preventable diagnoses within 30 days of starting outpatient chemotherapy[17]. Based on this measure, a study population at a comprehensive cancer center, including a large tertiary outpatient clinic, was assembled for risk prediction at 30, 180 and 365 days after chemotherapy initiation5. The OP-35 metric itself is used as a label for supervised learning and defines a positive event. For the SHD inputs, we use the original 760 features from Peterson et al.[5] extracted from the same EHR database, such as demographic, social, vital sign, procedural, diagnostic, medication, laboratory, health care utilisation, and cancer-specific data generated 180 days before the first date of chemotherapy. For a detailed description of how the patient cohort was extracted, the inclusion and exclusion criteria for the OP-35 metric, and a full list of features, please see the original paper[5].

Based on the above study population, we matched patients to their respective progress notes and the history and physical (H&P) notes from the EHR database (Epic Systems Corp). We removed notes of fewer than 100 words, as these were mainly erroneous entries, and notes of more than 5,000 words, as these often contained long copies of previous notes and laboratory analyses. We also removed history notes with mentions of clinical trial consents, as based on our review these were copy-paste texts. Finally, we extract and aggregate the most recent clinical notes (at most three) created 180 days before the patient started chemotherapy, same as in the SHD collection from Peterson et al.5. If a patient had no clinical records in the EHR database, they were removed from the study population.

The cohort was previously randomly divided into a training set (80%) and a test set (20%) for modelling, and we, therefore, kept exactly these patient sets (except the ones without any clinical notes) to obtain comparable results.

Model Development

Five different risk prediction models were compared in this study: Tabular LASSO, Language LASSO, Fusion LASSO, Language BERT and Fusion BERT.

Tabular LASSO. This model is a logistic regression with an l1-penalty (also known as Least Absolute Shrinkage and Selection Operator - LASSO). The inputs were the 760 structured health data points from Peterson et al.[5].

Language LASSO. The Language LASSO model is an l1-penalised logistic LASSO regression with manually generated inputs from the clinical notes. The notes were preprocessed as follows: First, we removed special characters and personal, organisational, date and time entities using SpaCy's[18] part of speech tagging. Then we tagged negated terms with a "not_" using SpaCy's negator library. We removed auxiliary words, adpositions, determiners, interjections and pronouns. Subsequently, we lemmatised[19] the remaining words. Finally, we followed the method of Marafino et al.[10] by filtering out the 2,000 most frequent terms[1] of all the notes and weighting these words using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The Language LASSO has 2,000 input features corresponding to the 2,000 most frequently occurring words.

Fusion LASSO. The Fusion LASSO is also a logistic regression LASSO model. This time it uses both, the tabular data and TF-IDF values, as input features. We combined them to inspect if data extracted from the clinical notes has added value to SHD.

Language BERT. This model is a deep learning-based Bidirectional Encoder Representation of Transformers[20,21] (BERT). This model does not require manual feature engineering and can consume clinical notes with little preprocessing. We used a pre-trained distilBERT[22] model as the encoding structure,

as it requires less computation than the more familiar BERT or ClinicalBERT[14] models[2]. DistilBERT and other transformer models are often computationally limited by the input size of the text (often referred to as token length in the deep learning literature). To avoid these GPU memory overflows, we decomposed the clinical notes into chunks of at most 25 sequences, each 256 tokens (1 token ≈ 1 word), and ran them through the neural network. We aggregated the outputs of the transformers (also called embeddings) by averaging over the embeddings of the corresponding clinical note. We connected the averaged embedding (representing a complete clinical note) linearly to a single output neuron, whose value is divided into four sections. A sigmoid function is applied to assign a probability to each of these values. These four slices represent the probability distribution of an ACU event within the time intervals emanating from the different ground truth labels (P (x ≤ 30d), P (30d < x ≤ 180d), P (180d < x ≤ 365d) and P (x > 365d)). Since a patient who experienced an ACU event within the first 30 days is also eligible for an event within 180 days and 365 days, we add the corresponding probabilities to get the original ground truth interpretation of an ACU within 30 days (P (x ≤ 30d)), 180 days (P (x ≤ 180d)), 365 days (P (x ≤ 365d)) and not within 365 days (P (x > 365d)).

Fusion BERT. The fusion BERT model is the same as the language BERT model, except that the corresponding SHD are concatenated with the output embedding. The newly-concatenated embedding was then linearly connected to the output neuron. Figure 1 shows an overview of the fusion BERT.

The regularisation hyperparameters of the LASSO models were determined by tenfold cross-validation grid search, while the hyperparameters of the two BERT models were determined by using 20% of the training data as validation data. While the LASSO models are trained on each time interval of the label (30d, 180d, and 365d) individually, the BERT models are trained on all the labels simultaneously. We applied a cumulative link loss[23] to train the neural network with backpropagation, to ensure the ordinal regression structure[24] (e.g avoid cases where P(x ≤ 30d) > P(x ≤ 180d), as in this use case any patient that had an ACU event within 30 days, subsequently is marked to have had an ACU within 180).

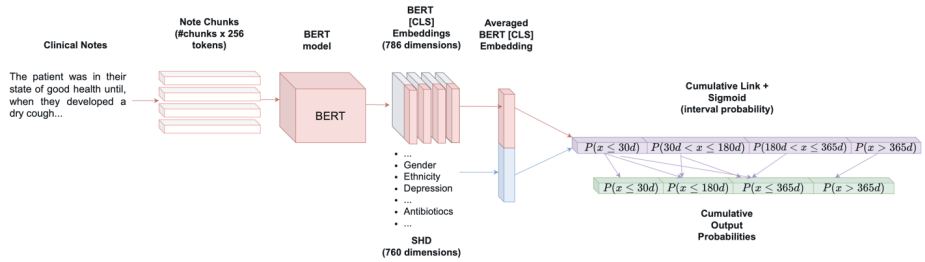


Figure 1. Overview of the Fusion BERT model. The Language BERT only contains the upper (red) encoder architecture before leading into the output probabilities (purple and green).

Model Evaluation

We first evaluated the models by their discriminative performance with the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC) and the negative log-likelihood (cross-entropy) with a 1'000-fold bootstrap to obtain 95% confidence intervals.

We reported the number of SHD used during risk prediction. We did this for the LASSO models by summing the number of non-zero model coefficients originating from SHD. For the Fusion BERT, we counted the number of connections of tabular features to the output neuron that have values less than 0.001, as the backpropagation algorithm is not optimized for feature selection, unlike the LASSO.

To assess the model calibration, calibration curves[25] were developed. Finally, we assessed the initial clinical utility of these four models through a Decision Curve Analysis (DCA)[26]. A DCA plots the net benefit across a range of decision thresholds and quantifies the number of true positives versus false positives. The curves for the prediction models were compared to two alternative clinical strategies: treat all (everyone is treated as if they will have an ACU event) and treat none (nobody is treated as if they will have an ACU event). To test the discriminatory power of the model in a setting similar to that in which it might be used at the point of care, the test cohort was stratified into high, medium

and low-risk groups based on the tertile of predicted risk. Kaplan-Meier[27] survival curves for OP-35 events are used to examine the separation between risk groups for language LASSO and language BERT on 180-day ACU risk prediction. In addition, the ten highest and lowest coefficients of the language LASSO model are presented. This helps us to determine the importance of certain keywords in the clinical notes.

Since Peterson et al.[5] have reported unfair algorithmic results for ACU prediction from structured data, we investigated whether language features might also be affected. We present and compare the empirical cumulative distributions of predicted risk score percentiles for subgroups to assess how the models predicted each subgroup's risk for OP-35 events. Specifically, we examine demographic values (i.e. race and insurance type) and tumour stage on the Language LASSO model for 180-day ACU risk prediction.

3.4 Results

A total of 6,938 patients were included in the study cohort compared to the original cohort, which included 8,439 patients. The mean age at chemotherapy initiation was 60.5 years (± 14.4 years), and 52.7% were female. A total of 936 patients (13.5%) met the primary criteria of having at least one OP-35 event within the first 30 days of starting chemotherapy, 2,202 (31.7%) within the first 180 days and 2,704 (39.0%) within the first year. The majority of patients in the cohort were white (n=3,804, 54.8%), followed by Asian patients (n=1,619, 23.3%), then other and unknown races (1,327, 19.1%) and least represented were black patients (n=188, 2.7%). The most common cancer type was breast cancer (n=1,321, 19.0%), gastrointestinal tumours (n=819, 11.8%), thoracic cancer (n=774, 11.2%) and lymphoma (n=700, 10.1%), which accounted for more than half of all data. ACU events occurred most frequently in lymphoma (30d: n=170, 18.2%; 180d: n=345, 15.7%; 365d: n=382, 14.1%) and least frequently in prostate cancer (30d: n=11, 1.2%; 180d: n=46, 2.1%; 365d: n=70, 2.6%) across all time periods.

Table 1. Information about the complete patient cohort (train and test set) eligible for the OP-35 metric for 30, 180, and 365 day prediction. "std" stands for standard deviation.

Patient Characteristic	Total Cohort (N=6,938)	OP-35 Events within 30 days (n=936, 13.5%)	OP-35 Events within 180 days (n=2,202, 31.7%)	OP-35 Events within 365 days (n=2,704, 39.0%)	No OP-35 Events within 365 days (n=4,234, 61.0%)
Age, mean ± std At diagnosis	58.7±14.3	57.2±15.3	57.7±15.1	57.9±15.0	59.2±13.9
At first chemotherapy	60.5±14.4	58.9±15.2	59.4±15.1	59.6±15.0	61.0±14.0
Sex, No. (%) Female	3,659 (52.7)	474 (50.6)	1,132 (51.4)	1,417 (52.4)	2,242 (53.0)
Race, No. (%) White	3,804 (54.8)	461 (49.3)	1,113 (50.5)	1,379 (51.0)	2,425 (57.3)
Asian	1,619 (23.3)	226 (24.1)	536 (24.3)	649 (24.0)	970 (22.9)
Black	188 (2.7)	42 (4.5)	88 (4.0)	100 (3.7)	88 (2.1)
Other or unknown	1,327 (19.1)	207 (22.1)	465 (21.1)	576 (21.3)	751 (17.7)
Ethnicity, No. (%) Non Hispanic/Latino	5,989 (86.3)	788 (84.2)	1,867 (84.8)	2,280 (84.3)	3,709 (87.6)
Hispanic or Latino	855 (12.3)	142 (15.2)	327 (14.9)	414 (15.3)	441 (10.4)
Cancer type, No. (%) Breast	1,321 (19.0)	113 (12.1)	275 (12.5)	346 (12.8)	975 (23.0)
Gastrointestinal	819 (11.8)	93 (9.9)	291 (13.2)	366 (13.5)	453 (10.7)
Thoracic	774 (11.2)	107 (11.4)	258 (11.7)	326 (12.1)	448 (10.6)
Lymphoma	700 (10.1)	170 (18.2)	345 (15.7)	382 (14.1)	318 (7.5)
Head and neck	658 (9.5)	90 (9.6)	208 (9.4)	238 (8.8)	420 (9.9)

Table 1. (*Continued*)

Patient Characteristic	Total Cohort (N=6,938)	OP-35 Events within 30 days (n=936, 13.5%)	OP-35 Events within 180 days (n=2,202, 31.7%)	OP-35 Events within 365 days (n=2,704, 39.0%)	No OP-35 Events within 365 days (n=4,234, 61.0%)
Cancer stage, No. (%) Stage I	1,099 (15.8)	123 (13.1)	281 (12.8)	338 (12.5)	761 (18.0)
Stage II	1,415 (20.4)	141 (15.1)	336 (15.3)	410 (15.2)	1005 (23.7)
Stage III	964 (13.9)	131 (14.0)	351 (15.9)	429 (15.9)	535 (12.6)
Stage IV	1,898 (27.4)	327 (34.9)	759 (34.5)	937 (34.7)	961 (22.7)
Unknown	1,562 (22.5)	214 (22.9)	475 (21.6)	590 (21.8)	972 (23.0)
Insurance, No. (%) Medicare	2,683 (38.7)	323 (34.5)	788 (35.8)	970 (35.9)	1,713 (40.5)
Private	2,450 (35.3)	328 (35.0)	747 (33.9)	898 (33.2)	1,552 (36.7)
Medicaid	599 (8.6)	130 (13.9)	258 (11.7)	307 (11.4)	292 (6.9)
Other or unknown	1,206 (17.4)	155 (16.6)	409 (18.6)	529 (19.6)	677 (16.0)

Table 2. Resulting metrics on the test set of the tabular, language and fusion LASSO models, as well as the language and fusion BERT, trained on 30, 180 and 365 days ACU prediction with the 95%-CI in the brackets. The best-performing metrics for every label type are marked in bold. We also display the number of SHD used for prediction in the third column, where "N/A" means that SHD was used for prediction.

Label	Model	No. SHD	AUROC	AUPRC	Cross-entropy
30	Tabular LASSO C=0.02	83	0.775 (0.757,0.792)	0.411 (0.373,0.447)	0.344 (0.329,0.358)
	Language LASSO C=0.03	N/A	0.726 (0.707,0.744)	0.294 (0.264,0.323)	0.363 (0.346,0.379)
	Fusion LASSO C=0.02	73	0.778 (0.760,0.795)	0.410 (0.372,0.447)	0.341 (0.326,0.356)
	Language BERT	N/A	0.710 (0.692,0.729)	0.259 (0.235,0.282)	0.435 (0.415,0.455)
	Fusion BERT	419	0.766 (0.749,0.784)	0.315 (0.286,0.343)	0.393 (0.377,0.406)
180	Tabular LASSO C=0.03	221	0.748 (0.735,0.762)	0.623 (0.600,0.647)	0.540 (0.527,0.552)
	Language LASSO C=0.02	N/A	0.730 (0.717,0.745)	0.577 (0.555,0.601)	0.558 (0.546,0.570)
	Fusion LASSO C=0.02	101	0.765 (0.752,0.779)	0.632 (0.610,0.655)	0.530 (0.517,0.543)
	Language BERT	N/A	0.702 (0.688,0.717)	0.543 (0.517,0.567)	0.625 (0.603,0.644)
	Fusion BERT	419	0.753 (0.741,0.767)	0.620 (0.597,0.644)	0.548 (0.536,0.558)
365	Tabular LASSO C=0.02	150	0.763 (0.752,0.775)	0.704 (0.685,0.724)	0.559 (0.549,0.569)
	Language LASSO C=0.02	N/A	0.732 (0.719,0.745)	0.639 (0.618,0.661)	0.585 (0.575,0.595)
	Fusion LASSO C=0.02	115	0.770 (0.759,0.782)	0.702 (0.683,0.722)	0.553 (0.541,0.563)
	Language BERT	N/A	0.709 (0.695,0.723)	0.617 (0.594,0.640)	0.666 (0.647,0.683)
	Fusion BERT	419	0.760 (0.748,0.774)	0.695 (0.675,0.714)	0.565 (0.554,0.575)

Most chemotherapy patients had a stage IV tumour (n=1,898, 27.4%), which was also most common in ACU events (30d: n=327, 34.9%; 180d: n=759, 34.5%; 365d n=937, 34.7%). The most common type of insurance in the cohort was Medicare (n=2,863, 38.7%) and private health insurance (n=2,450, 35.3%). The cohort characteristics are summarised in Table 1.

Model Performance

Table 2 lists the AUROC, AUPRC, and cross-entropy scores including the 95% confidence intervals of the five risk models for 30-day, 180-day and 365-day ACU prediction. For the 30 day acute care risk prediction, the Fusion LASSO model performs best on AUROC (0.778, 95%-CI: 0.760, 0.795) and cross-entropy (0.341, 95%-CI: 0.326, 0.356), using 73 SHD features. The highest AUPRC score has the Tabular LASSO (0.411, 95%-CI: 0.373, 0.447) compared to the event rate of 13.5%, using 83 tabular variables.

For 180-day ACU prediction, the Fusion LASSO model performs best in all metrics with 101 SHD features. The Language LASSO has a 0.730 (95%-CI: 0.717, 0.745) AUROC score and the Language BERT achieves 0.702 (95%-CI: 0.688, 0.717), both of them without using any structured data.

In the full-year ACU prediction, we observe that the Fusion LASSO scores again the highest C-stastic (0.770, 95%-CI:0.759, 0.782) and the lowest cross-entropy loss (0.553, 95%-CI:0.541, 0.563), using 115 tabular features, while the Tabular LASSO has the highest AUPRC score (0.704, 95%-CI:0.685, 0.724), using 150 SHD points. We show the flexible calibration curves for the 180-day models in Figure 2, where we observe a risk underestimation of the three LASSO models and underestimation of low risk patients and overestimation of high risk patients with the Language BERT model.

The Fusion BERT uses the most SHD points (419 tabular inputs) for all three label types to make predictions.

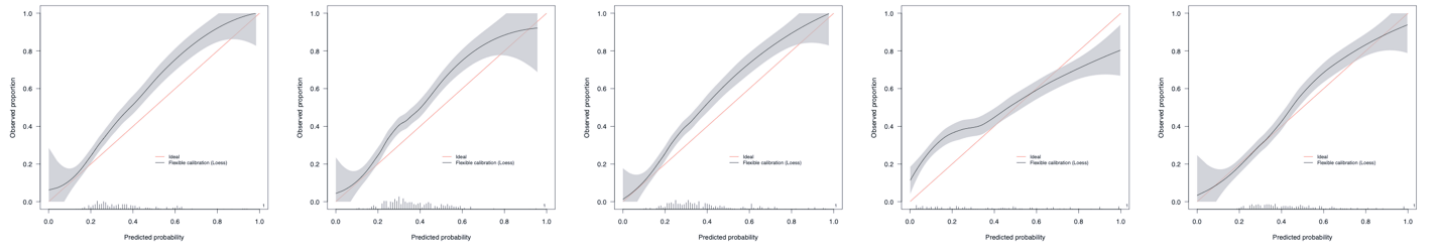


Figure 2. Calibration curves of the 180-day ACU risk prediction models. The red line indicates ideal calibration, while the black line is the flexible calibration with the 95% confidence interval, generated with the prob.cal.ci.2 function[25].

Exploration of Clinical Usage of Language Models

The Decision curve analysis for the 180-day ACU prediction showed that the net benefit of the Language-BERT model yields a negative benefit when the decision threshold for treatment is chosen above 0.6 (Figure 3) and less or equal net benefit than treating every patient with a threshold below 0.19. The other models, including the Language-LASSO model, have positive benefit values for decision thresholds below 0.7.

The Kaplan-Meier survival curves for OP-35 events showed good separation between risk groups (Figure 4, $p < 0.001$ for each group by log-rank test) for the two language-only models. By 180 days after the start of chemotherapy, 64 (13.9%) of the 462 low-risk patients in the language LASSO prediction had an OP-35 event and 76 (16.5%) in the language BERT prediction. On the other hand, 246 (53.2%) of the 462 high-risk patients had an event for the speech LASSO prediction and 238 (51.5%) for the language BERT prediction. Figure 5 shows the relative importance of the ten highest and lowest coefficients of the language LASSO model for the 180-day prediction. The words “Admission”, “Failure”, “Pain” and “Palliative” are among the ten highest coefficients, while “Breast”, “PSA (Prostate-specific antigen)”, “Nourished” and “Prostate” are among the ten lowest coefficients.

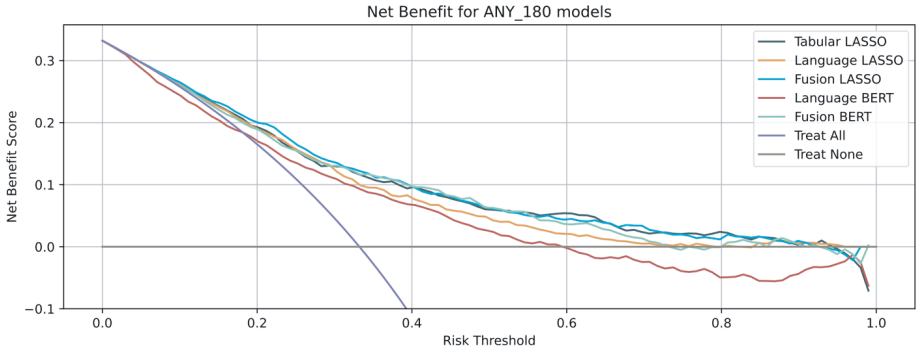
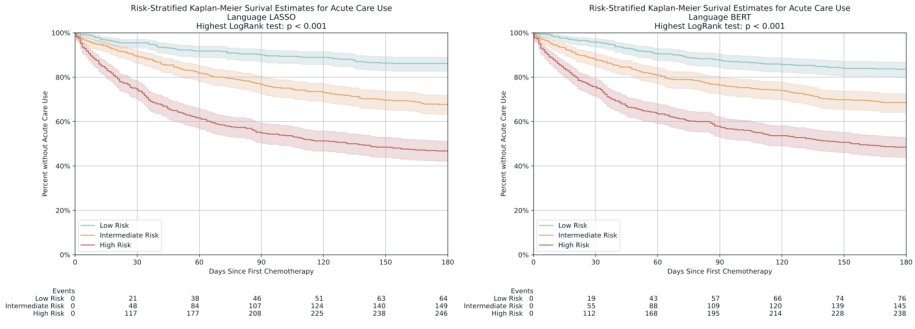


Figure 3. Net benefit curves of the tabular, language and fusion models. The purple curve indicates the benefit of all the patients treated, whereas the grey curve indicates the benefit is no patient is treated.



(a) Language LASSO (180d predictions) **(b)** Language BERT (180d predictions)

Figure 4. Kaplan-Meier curves for ACU events for patients in the test cohort stratified by predicted risk. The shaded area represents the 95%CIs.

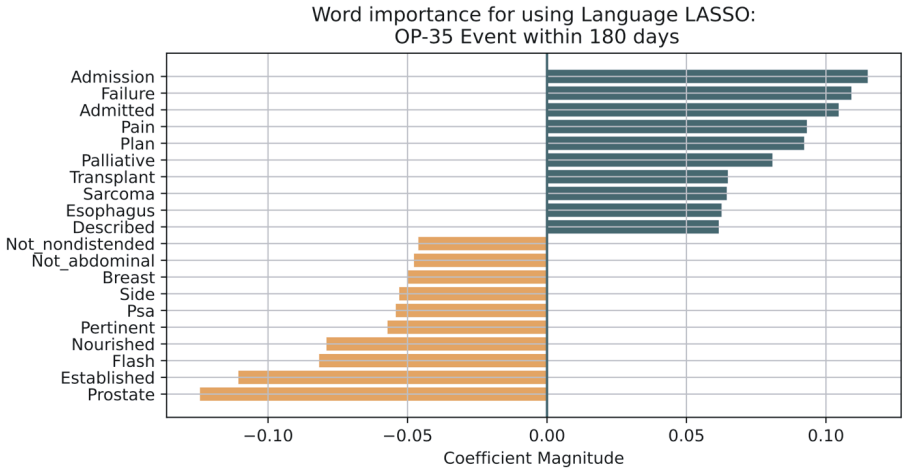
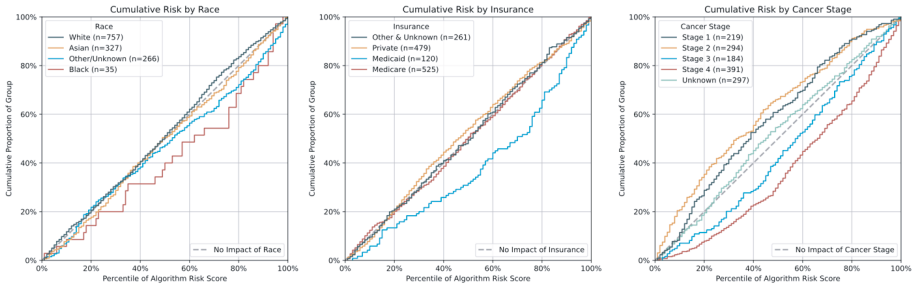


Figure 5. Coefficient magnitudes for the Language LASSO for 180-day ACU prediction, displaying the ten highest and ten lowest. The coefficients in this model are single words found in the clinical notes before the ACU event.

Sensitivity Analysis

Figure 6a shows that black patients are predicted to have a disproportionately higher risk than white, Asian or other-race patients. We note that the number of black patients is at least seven times smaller than that of non-black races. The cumulative risk by different insurance types is displayed in Figure 6b, where we note a risk overestimation of Medicaid patients.

In Figure 6c we see that the risk predictions for patients with stage III and IV tumours are overestimated, while the predictions for patients with stage I, II and unknown stage cancer are underestimated.



(a) Cumulative risk by race **(b)** Cumulative risk by insurance **(c)** Cumulative risk by cancer stage

Figure 6. Cumulative risk of the Language LASSO on 180 ACU prediction, stratified by race (a), insurance type (b) and over cancer stage (c).

3.5 Discussion

Identifying methods to improve external generalisability is a priority for the medical informatics community. This paper presented an analysis of NLP to identify patients at risk of seeking acute care for different time intervals, a method which is scalable across sites and required minimal data preprocessing. It presents a comparison of logistic regression LASSO using structured health data with manually generated language features and the combination of both modalities. In addition, these models are compared to deep learning-based transformers using either full-text clinical records only or in combination with structured

health data. Results demonstrate tabular LASSO outperforming language LASSO and language BERT at all time intervals. Nevertheless, we point out that both language-based methods achieve good discriminative performance (AUROC ≥ 0.7) even without the use of tabular health data, especially for 180-day risk prediction. On the other hand, combining the two input modalities (clinical notes and SHD) did not yield significant improvements over using tabular data alone. Regarding the selection model, results show that the popular BERT-based classifier does not outperform l1-penalised logistic regression with TF-IDF features and the fusion of both input modalities. This is likely due to the aggregation of chunks of the lengthy clinical documents into a single output probability, which makes deep learning training difficult.

Our study has several strengths; first, we show that NLP methods can be used instead of high-dimensional SHD to identify chemotherapy patients at risk for acute treatment. Secondly, our method optimises considerably training as we developed a transformer-based model that is trained simultaneously on multiple risk intervals in the form of nested ordinal regression, so that the computationally intensive training of the model for the different labels is not required.

Three main clinical implications can be drawn from this study. First, ACU risk prediction models for chemotherapy patients perform well using free-text data from the last (at most three) H&P and progress notes from physicians before chemotherapy. This implies that NLP methods could be easily implemented across sites and/or facilities as they only require access to written medical notes without the need for re-structuring or mapping of structured data, potentially saving costs in feature collection. Second, we show that LASSO coefficients can be used by clinicians to understand relative word meaning when making a prediction. BERT models lack this type of interpretive mechanism that allows clinicians to build confidence in the model. Third, in the sensitivity analysis, we find that certain groups may be subject to risk bias, despite not using demographic values explicitly as inputs. Although this needs further investigation, the results suggest that clinicians should be aware of these subgroup differences when interpreting the results of ML models.

Our study has limitations. First, more sophisticated transformer language models can handle longer notes[28] or also cross-modality[29]. However, these techniques require significant computation, which may not be feasible at most institutes. Second, our models have been validated only on one dataset for risk prediction in acute care. It would be beneficial to test the language models in a variety of medical problems. Finally, the work relies on data from a single academic cancer center, which may be not generalisable to other populations. In summary, this study demonstrates the utility of using free-text data to identify patients at risk of needing acute care once they have started chemotherapy. It is an alternative to structured health data, which may require significant preprocessing and may not be generalisable across settings. We show that the Language LASSO is a suitable model, especially for 180-day prediction, and that it is still interpretable. This work advances our knowledge on risk prediction models and provides an alternative for cross site generalisability. Hospital systems may consider these techniques to validate risk models.

Data and Code Availability

Under the terms of the data sharing agreement for this study, we cannot share source data directly. Requests for anonymous patient-level data can be made directly to the authors. All experiments were implemented in Python using the library SciKit-learn[30] for the metrics and the classical logistic regression model, and PyTorch[31] for the transformer models. We used R to create the calibration plots. The code for the logistic LASSOs and our analysis is available on www.github.com/su-boussard-lab/nlp-for-acu, while the code for the BERT model and training is available on www.github.com/su-boussard-lab/bert-for-acu.

References

1. Brooks GA, Li L, Uno H, Hassett MJ, Landon BE, Schrag D. Acute hospital care is the chief driver of regional spending variation in Medicare patients with advanced cancer. Health Aff (Millwood). 2014 Oct;33(10):1793-800.
2. Yabroff KR, Lamont EB, Mariotto A, Warren JL, Topor M, Meekins A, et al. Cost of care for elderly cancer patients in the United States. J Natl Cancer Inst. 2008 May;100(9):630-41.
3. Adelson KB, Dest V, Velji S, Lisitano R, Lilenbaum R. Emergency department (ED) utilization and hospital admission rates among oncology patients at a large academic center and the need for improved urgent care access. Journal of Clinical Oncology. 2014;32(30_suppl):19-9. PMID: 28141471. Available from: https://doi.org/10.1200/jco.2014.32.30_suppl.19.
4. Uno H, Jacobson JO, Schrag D. Clinician assessment of potentially avoidable hospitalization in patients with cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32 30_suppl:4.
5. Peterson DJ, Ostberg NP, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions. JCO Clinical Cancer Informatics. 2021;(5):1106-26. PMID: 34752139. Available from: https://doi.org/10.1200/CCI.21.00116.
6. Brooks GA, Kansagra AJ, Rao SR, Weitzman JI, Linden EA, Jacobson JO. A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy. JAMA Oncology. 2015 07;1(4):441-7. Available from: https://doi.org/10.1001/jamaoncol.2015.0828.
7. Grant RC, Moineddin R, Yao Z, Powis M, Kukreti V, Krzyzanowska MK. Development and validation of a score to predict acute care use after initiation of systemic therapy for cancer. JAMA Network Open. 2019;2.
8. Daly RM, Gorenshteyn D, Gazit L, Sokolowski S, Nicholas K, Perry C, et al. A framework for building a clinically relevant risk model. Journal of Clinical Oncology. 2019.
9. Office of the National Coordinator for Health Information Technology: National trends in hospital and physician adoption of electronic health records; 2022. https://www.healthit.gov/data/quickstats/ national-trends-hospital-and-physician-adoption-electronic-health-records.
10. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA Netw Open. 2018 Dec;1(8):e185097.
11. Marafino BJ, Davies JM, Bardach NS, Dean ML, Dudley RA. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J Am Med Inform Assoc. 2014 Sep;21(5):871-5.
12. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. J Pers Med. 2020 Dec;10(4):286.
13. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. 2020 Mar;27(3):457-70.

14. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72-8. Available from: https: //www.aclweb.org/anthology/W19-1909.
15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:190405342. 2019.
16. Sarraju A, Coquet J, Zammit A, Chan A, Ngo S, Hernandez-Boussard T, et al. Using deep learning-based natural language processing to identify reasons for statin nonuse in patients with atherosclerotic cardiovascular disease. Communications Medicine. 2022;2.
17. 2019 chemotherapy measure facts admissions and emergency department (ED) visits for patients receiving outpatient chemotherapy hospital outpatient quality reporting (OQR) program (OP-35);. https:// qualitynet.cms.gov/files/5dcc6762a3e7610023518e23?filename=-CY21_OQRChemoMsr_FactSheet.pdf .
18. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. In: AMIA Annual Symposium Proceedings 2021; (in press, n.d.). Available from: http://arxiv.org/abs/2106.07799.
19. Miller GA. WordNet: A Lexical Database for English. Commun ACM. 1992;38:39-41.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all You need; 2017. Available from: https://arxiv.org/pdf/1706.03762.pdf.
21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. 2019;abs/1810.04805.
22. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv; 2019.
23. Pedregosa F, Bach FR, Gramfort A. On the consistency of ordinal regression methods. J Mach Learn Res. 2017;18:55:1-55:35.
24. Rosenthal E. Spacecutter: ordinal regression models in pytorch; 2018. Available from: https:// www. ethanrosenthal.com/2018/12/06/spacecutter-ordinal-regression.
25. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016 Jun;74:167-76.
26. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019;3.
27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958 Jun;53(282):457.
28. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv:200405150. 2020.
29. Khadanga S, Aggarwal K, Joty SR, Srivastava J. Using clinical notes with time series data for ICU management. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics. 2019;6432-6437.
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. Journal of machine learning research. 2011;12(Oct):2825-30.

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024-35.

Chapter 4

Analyzing Patient Experiences using Natural Language Processing: Development und Validation of the Artificial Intelligence Patient Reported Experience Measure (AI-PREM)

Marieke M. van Buchem, Olaf M. Neve,
Ilse M.J. Kant, Ewout W. Steyerberg,
Hileen Boosman & Erik F. Hensen

BMC Medical Informatics & Decision Making,
15 July 2022

4.1. Abstract

Background

Evaluating patients' experiences is essential when incorporating the patients' perspective in improving healthcare. Experiences are mainly collected using closed-ended questions, although the value of open-ended questions is widely recognized. Natural language processing (NLP) can automate the analysis of open-ended questions for an efficient approach to patient-centeredness.

Methods

We developed the Artificial Intelligence Patient-Reported Experience Measures (AI-PREM) tool, consisting of a new, open-ended questionnaire, an NLP pipeline to analyze the answers using sentiment analysis and topic modeling, and a visualization to guide physicians through the results. The questionnaire and NLP pipeline were iteratively developed and validated in a clinical context.

Results

The final AI-PREM consisted of five open-ended questions about the provided information, personal approach, collaboration between healthcare professionals, organization of care, and other experiences. The AI-PREM was sent to 867 vestibular schwannoma patients, 534 of which responded. The sentiment analysis model attained an F1 score of 0.97 for positive texts and 0.63 for negative texts. There was a 90% overlap between automatically and manually extracted topics. The visualization was hierarchically structured into three stages: the sentiment per question, the topics per sentiment and question, and the original patient responses per topic.

Conclusions

The AI-PREM tool is a comprehensive method that combines a validated, open-ended questionnaire with a well-performing NLP pipeline and visualization. Thematically organizing and quantifying patient feedback reduces the time invested by healthcare professionals to evaluate and prioritize patient experiences without being confined to the limited answer options of closed-ended questions.

4.2 Background

Patient-centeredness is an essential fundament for providing high-quality care[1,2]. Insight into the patient-centeredness of care is obtained by evaluating patient experiences, typically using Patient-Reported Experience Measures (PREMs). Most PREMs include a combination of closed- and open-ended questions. When presented with both, healthcare professionals tend to value the answers to open-ended questions most[3]. These answers can be used to identify new points of interest ('topics') and provide context to closed-ended questions[3,4]. Although the value of open-ended questions is widely recognized, patients' free-text answers remain underutilized in clinical practice. One of the key challenges lies in the time needed for analysis. The answers to open-ended questions are often manually analyzed, which is laborious and time-consuming[3], especially in larger groups of patients.

Several studies aim to automate the analysis of free-text patient experience data to inform quality improvements, showing promising results[5,6,7,8,9,10,11,12,13,14,15]. Most of these studies concentrate on publicly available social media or forum data, usually focused on reviewing hospitals or physicians[5,6,7,8,9]. Current approaches include the use of artificial intelligence (AI) methods such as machine learning and natural language processing (NLP). A few studies successfully applied NLP techniques to routinely collected PREM questionnaires of patients[10,11,12,13,14,15]. Most of these studies use supervised methods; for example, topic classification is used to classify data into predefined, manually extracted topics[5,7,11,13,15]. Although some of these methods perform well, supervised methods lack the capability of finding new or unexpected topics. Moreover, regular manual labeling is time-consuming and, therefore, not suited to decrease the current burden of reading through the patients' answers[10]. Using unsupervised methods such as topic modeling can overcome these limitations. Two studies have compared supervised topic classification to unsupervised topic modeling and concluded that topic modeling leads to topics similar in quality[7,15].

Current open-ended questions are often unsuitable for automatic analysis as they were not developed for this purpose[10,11]. An example is a questionnaire consisting of the questions 'What did we do well?' and 'What could we improve?'. Previous work shows that answers to both questions can be positive and negative, complicating automated sentiment analysis[10,11]. One study created a new, open-ended questionnaire suitable for analysis with NLP[16], focusing on patient-reported outcomes instead of experiences. They concluded that adding open-ended questions leads to richer, more in-depth information, and analysis with NLP makes it feasible to use in clinical practice.

The aim of this study is to harness the value of free-text patient experiences, using NLP methods that have the flexibility to find new topics in a complex, fast-changing environment. Our approach is to develop and validate a method for collecting and analyzing open-ended PREMs that could be incorporated into clinical practice. This objective contains three sub-objectives:

1. Develop and validate an open-ended generic PREM questionnaire;
2. Develop and validate an NLP pipeline to automatically analyze the open-ended PREM;
3. Develop a visualization that supports healthcare professionals in identifying quality improvements from the results.

4.3 Methods

We devised a method that included a new, open-ended questionnaire, an NLP pipeline to analyze the questionnaire, and a visualization of the output of the NLP pipeline (Figure 1). This project was organized in a development phase and a validation phase. The development phase started with developing a new questionnaire, the Artificial Intelligence Patient-Reported Experience Measure (AI-PREM).

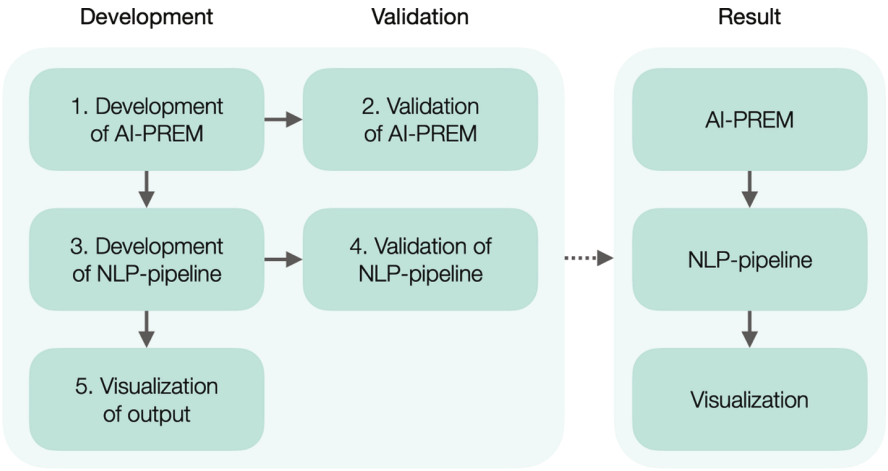


Figure 1: Overview of the different tasks and phases.

Development of the AI-PREM (Figure 1, step 1)

The AI-PREM was developed iteratively with patients from the vestibular schwannoma care pathway in the Leiden University Medical Center (LUMC) (Box 1). We used the following criteria: (1) Open-ended questions; (2) Phrasing suitable for analysis with NLP; (3) Generic questions, therefore not containing disease-, department-, or center-specific questions; (4) Accessible in terms of length and language. The Picker principles of patient-centered care[17] were the basis for the questionnaire. The development process started with questions about all eight Picker principles, asking patients about experiences with the accessibility of care, continuity of care, involvement of family, emotional support, information provision, physical needs, and involvement in decisions. Each question included one subject and did not contain a sentiment, to decrease the variability of patients' answers. For example, instead of asking 'What could be improved in the organization of care?' the question stated 'How was the organization of care?'. These questions were evaluated and finetuned in a group of patients.

Box 1: Description of the vestibular schwannoma care pathway in the LUMC.

> Vestibular schwannomas are benign intracranial tumors, with a heterogeneous clinical presentation: it may present as a small, slow growing, and asymptomatic tumor, but also as large, faster growing, and potentially fatal disease. Patients typically present with symptoms of hearing loss, loss of balance and vertigo, but may also suffer from facial numbness, facial paralysis, or elevated intracranial pressure. In non-progressive tumors, active surveillance with MRI is usually the management option of choice. In progressive tumors, surgery or radiotherapy is performed to prevent future complications. After an active intervention, prolonged active surveillance ensues in these patients too, in order to identify possible recurrences. The LUMC is an expert referral center for vestibular schwannoma in the Netherlands. The care is organized in an integrated practice unit including all specialties involved in the diagnosis and treatment (i.e., neurosurgery, otorhinolaryngology, radiology and radiation oncology).

Patients who participated in a survey study in 2014 were re-approached for participation in the AI-PREM project between May and September 2020[18]. Patients that agreed to participate provided their written informed consent. All patients were diagnosed with unilateral vestibular schwannoma between 2003 and 2014. Patients with bilateral vestibular schwannoma, other skull base pathologies, or insufficient proficiency in the Dutch language to complete the questionnaires were excluded. In addition to the AI-PREM, patients were also asked to fill out a validated structured patient experience questionnaire, the patient experience monitor (PEM), for comparison[1]. Patients first filled out the AI-PREM to ensure they were not biased towards the topics assessed in the PEM. The questionnaires were sent out either by e-mail using Castor software or hard copy by mail. These hard copies were verbatim digitalized manually.

Validation of the AI-PREM (Figure 1, step 2)

To validate the AI-PREM questionnaire, we used the COSMIN reporting guideline for studies on the measurement properties of patient-reported outcome measures[19]. Although this guideline is aimed at structured questionnaires about

patient outcomes, most parts can be applied to unstructured patient experience questionnaires. The COSMIN guideline investigates the content validity of questionnaires by looking at the questions' relevance, comprehensiveness, and comprehensibility. We examined the content validity of the AI-PREM by comparing AI-PREM questions to similar questions from the PEM. First, a sentiment analysis (as described in the Sentiment analysis section under 'Development of the NLP pipeline') was performed, labeling a text as positive or negative feedback. We hypothesized that patients who were negative about certain aspects of care in the AI-PREM would also give lower scores on the matched PEM questions and vice versa (scores range from one to ten, where one is the lowest and ten is the highest). Therefore, we defined 'positive' and 'negative' comments per AI-PREM question based on the sentiment analysis. Per AI-PREM question, we took the matched PEM questions and calculated the average score for the 'positive' and 'negative' groups. Using a t-test for independent samples, we compared the average scores between the 'positive' and 'negative' groups.

Development of the NLP pipeline (Figure 1, step 3)

The pipeline as described by Cammel et al. was taken as a starting point[10]. The pipeline includes sentiment analysis, preprocessing, and topic modeling. We combine a supervised (sentiment analysis) and unsupervised (topic modeling) approach. We use a supervised approach for the sentiment analysis because the categories for this task (positive, neutral, negative) will not change over time, in contrast to the topics that patients mention. The pipeline was developed in an iterative process by a team of data scientists, researchers, and clinicians of the vestibular schwannoma IPU, to fulfill the following pre-set requirements:

- Interpretable: The end-user should be able to distill from the output what patients experience as positive and negative.
- Actionable: The output should be specific enough to lead to concrete action points.
- Complete: The number of texts that cannot be assigned to a topic should be as small as possible.

Once the output met all the requirements according to the development team, the validation phase started.

Sentiment analysis

We finetuned a pretrained, multilingual BERT model for two binary classification tasks for sentiment analysis. The first binary classification task classified answers as negative or non-negative; the second task classified the non-negative answers as positive or neutral. To train these two sentiment analysis models, one annotator (MvB) manually labeled 75% of the collected data as 'negative', 'positive', or 'neutral'. A second annotator (ON) labeled 1/3rd of this data (25% of the collected data), which was used to calculate the inter-annotator agreement (percentage of datapoints that the annotators agreed on). Annotators labeled an answer as 'negative' if it described a topic or situation that the patient was dissatisfied with (e.g., 'I had to wait for a long time'). If a non-negative answer described a topic or situation that the patient was satisfied with, it was labeled as 'positive' (e.g., 'the personnel was very friendly'). All answers that described a topic or situation that was neither positive nor negative were labeled as 'neutral' (e.g., 'first I was treated at hospital number 1, then I was referred to hospital number 2'). The two sentiment analysis models were trained on a random sample of 80% and validated on the other 20% of labeled data, using the default parameters of the Transformers implementation of the BERT model for Sequence Classification[20].

Preprocessing

After the sentiment analysis, the data were preprocessed. We tokenized words and corrected the spelling using the Peter Norvig algorithm[21] and the CyHunSpell Python package[22]. Subsequently, words were lemmatized, and all non-informative words (stopwords, words with less than three letters, and all words except verbs, adverbs, nouns, and adjectives) were removed using the Stanza Python package[23]. Finally, all n-grams ranging from one to three were vectorized using term frequency-inverse document frequency (TF-IDF).

Topic modeling

We used topic modeling, specifically Non-negative Matrix Factorization (NMF), to identify the most important topics from the patients' answers to the AI-PREM, as described by Cammel et al.[10]. NMF was chosen over Latent Dirichlet Allocation because patients' answers tend to be very short and NMF is better able to deal with short answers. A separate topic model was created per sentiment (positive or negative) and per question. For each topic model, the optimal number of topics was chosen by creating several topic models with topics ranging from 2 to 15 and calculating the coherence score within every topic. The coherence score was calculated using the semantic similarity of words within a topic, based on a Dutch Word2Vec model[24,25,26], to account for exact matches and synonymous words. The topic model with the highest coherence metric was chosen as the best fitting model for that specific category.

Validation of the NLP pipeline (Figure 1, step 4)

We performed different validation steps to evaluate the performance of the NLP pipeline. (1) We assessed whether the automatically defined topics were representative of the texts they described. (2) We evaluated whether the NLP pipeline extracted topics similar to human-extracted topics.

Representativeness of the data

We randomly sampled the answers to the AI-PREM and performed manual evaluations of these answers by clinical experts. One clinician (ON) assessed a sample of the texts within the different categories (e.g., positive answers about information, negative answers about the organization of care). Per category, 20% of the answers per topic were analyzed, with a minimum of 10 texts. Some topics included less than ten texts; the clinician evaluated all texts for these topics. For every text within the sample, the clinician decided if it fit within the assigned topic. This analysis resulted in a percentage showing how representative the different topics were for the answers within that topic. A researcher (MvB) went through the same validation process to calculate the inter-annotator agreement.

Topic model versus human comparison

To investigate the performance of the topic model compared to human analysis, two clinical experts (a physician and a nurse practitioner) from the vestibular schwannoma care pathway read the answers to the AI-PREM from a sample of 50 patients, as data saturation was reached. A qualitative approach was used to identify topics within these texts. After reading, the experts decided on a few topics per question that summarized patients' answers in a consensus meeting. Two researchers (MvB and ON) compared these manually selected topics to the automatically selected topics from the NLP pipeline. Because the human analysis consisted of a sample of 50 questionnaires (and not all), we did not try to match exact words but matched on topic level. The proportion of manually identified topics that could be matched to an automatically identified topic was subsequently calculated.

Visualization of the output (Figure 1, step 5)

To stimulate the use of the AI-PREM tool in clinical practice, we co-created a mock-up of a potential visualization. We held three feedback sessions with a group of physicians, nurse practitioners, and implementation managers and iteratively updated the visualization based on their feedback and pre-set requirements. The requirements for the visualization were:

- Applicability within the end-users current workflow;
- Presentation of an overview of the output at a glance;
- Ability to get more context without going through all the individual questionnaires.

4.4 Results

Development of the AI-PREM

During six iterations, the initial questions were finetuned. The most significant changes made during these iterations were reducing the number of questions and simplifying the sometimes abstract Picker principles. The comprehensibility improved by using only level B1 words of the Common European Framework of

Reference for Languages [27]. Furthermore, patients preferred to have some examples of what was meant by the different aspects. The Picker institute provides some examples, which we added to each question. This led to the following questions:

- Q1: How was the provided information? Think of: the prognosis, possible tests, and treatment(s)
- Q2: How was the personal approach? Think of: shared decision making, listening to your preferences, emotional support
- Q3: How was the collaboration between healthcare professionals? Think of: no varying advice or having to tell your story multiple times, contact with your family doctor or other hospitals
- Q4: How was the organization of care? Think of: making appointments, combining appointments on one day, availability by phone
- Q5: What else would you like to share about your experience?

In total, 536 out of 867 vestibular schwannoma patients filled out the AI-PREM and PEM questionnaires, resulting in a response rate of 62%. Two patients were excluded because their diagnosis changed from vestibular schwannoma to meningioma, requiring treatment in another care pathway. This resulted in 534 sets of questionnaires. The median length of patients' answers was two words, with an interquartile range of 1 to 11 words. The maximum length was 192 words.

Validation of the AI-PREM

Using the Picker principles as a basis, the AI-PREM adhered to the relevance and comprehensibility criteria from the COSMIN reporting guideline. The comprehensibility criterium was further substantiated by including patients in the development of the AI-PREM. The results of validating the last criterium, comprehensiveness, are shown in Table 1. Where Q1-3 showed a significant difference in PEM scores between positive and negative answers, Q4 did not. No PEM questions were matched to Q5 ('What else would you like to share about your experience?'), so we did not validate this question.

Table 1: Overview of the number of AI-PREM responses per sentiment.

Questions	*Number of patients N (%)*	*Average PEM scores of matched questions, ranging from 1 to 10 μ ± sd*
Q1 – Positive *Negative*	359 (67.2%) 26 (4.9%)	9.7 ± 0.9 8.1 ± 2.4**
Q2 – Positive *Negative*	360 (67.4%) 31 (5.8%)	9.7 ± 0.7 7.7 ± 2.6**
Q3 – Positive *Negative*	325 (60.9%) 40 (7.5%)	9.6 ± 1.1 8.3 ± 1.8*
Q4 – Positive *Negative*	343 (64.2%) 39 (7.3%)	6.9 ± 1.7 6.4 ± 2.0
Q5 – Positive *Negative*	121 (22.7%) 35 (6.6%)	

The neutral responses are left out. Per category (question and sentiment), the average scores to the PEM questions that matched the AI-PREM questions are shown. P-value for the t-test for independent samples: * = $p < 0.001$, ** = $p < 0.0001$. AI-PREM: artificial intelligence patient reported experience measure. PEM: patient experience monitor. Q: question. Sd: standard deviation.

Development of the NLP pipeline

We made several improvements to the pipeline during the iterative development process (Box 2). The final NLP pipeline contained a sentiment analysis model consisting of a negative and positive sentiment classifier and a topic modeling module (Figure 2).

Box 2: Most important improvements that were made during the iterative development process.

- To first perform a sentiment analysis and then create a separate topic model per sentiment and per question, instead of creating one topic model for both sentiments. This led to more specific topics, from which points of improvements could be derived more easily, increasing the interpretability and actionability
- To not only include the negative feedback topics but also the positive ones, in order to obtain more balanced information. This was found to be essential in selecting and prioritizing points of improvement. In addition, the positive topics were seen as motivators for the healthcare team

- To go from a fixed number of topics to an adaptive approach that automatically chooses the optimal number of topics per subject. This increased the completeness
- To add a quantitative dimension to the qualitative output of the topic model, in order to help prioritize aspects of care that need the most attention
- To include n-grams up to three instead of just using 1 g. This increased the interpretability and actionability of the topics

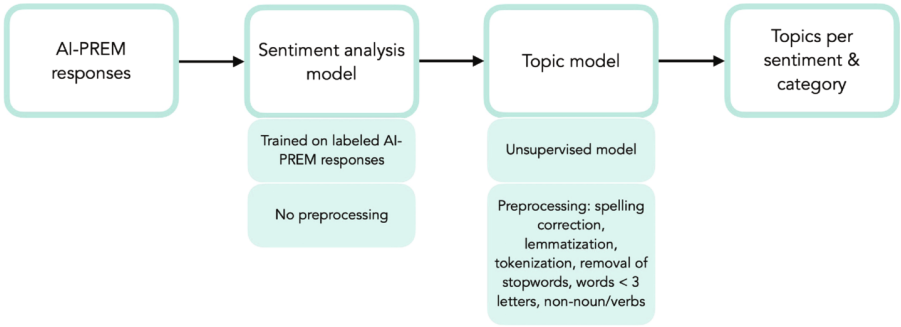


Figure 2: Overview of the input, models, and output of the AI-PREM tool.

Sentiment analysis

The inter-annotator agreement was 91.9%. The precision and recall for the negative sentiment model were 0.78 and 0.53, respectively, with an F1 score of 0.63. The precision, recall, and F1 score for the positive sentiment model were all 0.97.

Topic modeling

The number of topics per category ranged from two to six. 2.8% of texts could not be assigned to a topic. Only the ten n-grams with the highest TF-IDF score per topic were extracted to increase the interpretability of the topics. These n-grams were sorted based on the number of words, with the highest number of words shown first. We deduplicated this list of words to ensure that the final list of descriptors would not contain both 'went very well' and 'went well'. Finally, the first five words of this sorted, deduplicated list were shown to the end-user (Figure 3). See Additional file 1 for all the topics per category.

Q5: What else would you like to share about your experience?

Positive topics	Amount
Only good, good experience, experience good, aftercare good, very good	105
Treatment aftercare, only positive, only good, good experience, everyhing fine	8

Negative topics	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14
Long wait, result scan, wait result, long ago, surgery confess	21

Figure 3: Topic model for Q5.

Validation of the NLP pipeline

The overall percentage of representative texts was 80.9%, with 90.1% for the positive texts and 72.0% for the negative texts (Table 2). The inter-annotator agreement was 94.4% for positive texts, 80.5% for negative ones, and 90.4% overall. The clinical experts extracted 20 topics: 14 for the positive and 6 for the negative texts. All negative topics and 12 of 14 positive topics could be matched to the automatically extracted topics, leading to a 90% overlap between human topics and automatically extracted topics .

Visualization of the output

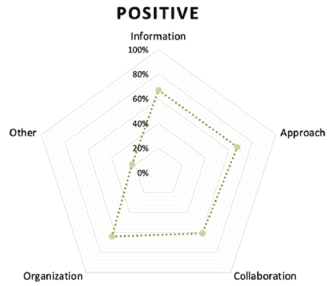
The end-users preferred the spider plot over other visualizations in the feedback session, such as a bar plot or tornado graph. The final visualization included a mock-up with three stages (Figure 4).

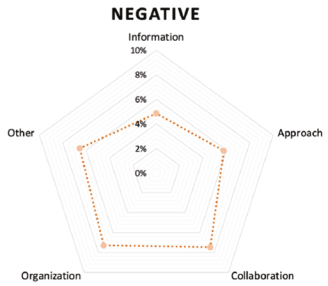
Table 2: Representativeness of the different topic models per category.

Question	Positive categories in total	Per topic	Negative categories in total	Per topic
Q1	94.4% (n = 72)	T1: 100% (n = 36) T2: 88.9% (n = 36)	55.6% (n = 18)	T1: 60% (n = 10) T2: 50% (n = 8)
Q2	93.3% (n = 75)	T1: 97.1% (n = 35) T2: 100% (n = 10) T3: 85% (n = 20) T4: 90% (n = 10)	71% (n = 31)	T1: 100% (n = 3) T2: 100% (n = 3) T3: 83.3% (n = 6) T4: 100% (n = 3) T5: 75% (n = 4) T6: 28.6% (n = 7) T7: 60% (n = 5)
Q3	98.4% (n = 63)	T1: 100% (n = 43) T2: 95% (n = 20)	76.9% (n = 39)	T1: 100% (n = 4) T2: 33.3% (n = 3) T3: 85.7% (n = 7) T4: 100% (n = 5) T5: 66.7% (n = 3) T6: 77.8% (n = 9) T7: 62.5% (n = 8)
Q4	100% (n = 65)	T1: 100% (n = 41) T2: 100% (n = 12) T3: 100% (n = 12)	86.7% (n = 15)	T1: 100% (n = 5) T2: 80% (n = 10)
Q5	86.2% (n = 29)	T1: 85.7% (n = 21) T2: 87.5% (n = 8)	55.5% (n = 20)	T1: 50% (n = 10) T2: 60% (n = 10)

Representativeness is defined as the number of texts within a certain topic that fit the description of the topic. The percentage is calculated by dividing the texts that fit the description of the topic by the total number of texts within the topic. Q: AI-PREM question. T: automatically extracted topic.

(a)





What else would you like to share about your experiences?

Positive clusters	Amount
Only good, good experience, experience good, aftercare good, very good	105
Treatment aftercare, only positive, only good, good experience, everyhing fine	8

Negative clusters	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14
Long wait, result scan, wait result, long ago, surgery confess	21

(b)

What else would you like to share about your experiences?

Positive topic 1	Amount
Only good, good experience, experience good, aftercare good, very good	105

Negative topic 1	Amount
Aftercare good, aftercare deal with, deal with new, situation well, new situation	14

Original patient responses

The interaction with all the staff was very good, they really listened to me
The contact with the nurse and doctor was very good
Good support by the nurse
Good
Good collaboration between the other hospital and the LUMC. It was easy to book a new scanning appointment and the information was clear.
I can only be positive about the treatment and the aftercare.

Aftercare could be improved
Aftercare could be slightly better
Exceptionally difficult to make the right decision. Thoroughly delved into the problem. Eventually chose the direct approach in the form of Cyberknife radiation.
Aftercare (dealing with new situation) could be better
Aftercare female eye doctor was so-so

Figure 4: a Stage 1: the spider plot showing the percentage of positive and negative texts per question. Stage 2: once the end-user clicks on one of the questions, the automatically extracted topics are shown. The positive topics are shown on the left and the negative topics on the right. b Stage 3: if the end-user wants to dive into one of the topics, they can click on that topic and read the actual patient answers that belong to that topic. In this example, the end-user is looking at the topics within the 'Other' category and has clicked on positive topic 1 and negative topic 1.

4.5 Discussion

This study describes the development and validation of a comprehensive tool for surveying the patient experience that can automatically produce actionable information. The tool consists of an open-ended, validated patient experience questionnaire suitable for qualitative and quantitative analysis with natural language processing (NLP), a well-performing NLP pipeline to analyze the answers to the questionnaire automatically, and a visualization that supports healthcare professionals in defining quality improvements from the results.

A critical aspect of our study is that we created and validated a new questionnaire consisting of only open-ended questions. One other study developed a new, open-ended questionnaire suitable for analysis with NLP, but they focused on patient outcomes instead of experiences[16]. Unique in our study is that we compared the AI-PREM with a 'gold standard' PREM, the patient experience monitor (PEM). Overall, three out of four open-ended questions of the AI-PREM seem to capture sentiments similar to the PEM. The lack of a significant correlation for the fourth question, asking about the organization of care, might be explained because this question had the lowest average PEM score and the smallest range.

Our NLP pipeline combines sentiment analysis with topic modeling while also making it possible to go back to individual patients' original responses per topic. This hierarchical structure allows healthcare professionals to scan the sentiment analysis for a high-level view or dive into the different topics and texts to define quality improvements. Physicians can use the quantitative data to review the results at a glance and prioritize the various topics, while the qualitative data allows them to put the topics into context and define concrete points of action.

Unlike most studies[5,7,11,13,15], we chose an unsupervised topic modeling approach due to its flexibility in finding new and unexpected topics[3,10]. One example that highlights the benefit of this approach is the topic describing the negative sentiment patients had about how long they had to wait for the scan results. This topic is not included in structured questionnaires and is very specific to this care pathway. Furthermore, the differing number of topics per question

shows the ability of this method to adapt to the data at hand. Methods sensitive to changing topics in patients' experiences are essential in the constantly changing healthcare environment.

We finetuned a pretrained multilingual BERT model on our data for the current sentiment analysis. Because the questionnaire and answers were in the Dutch language, there was limited choice in off-the-shelf sentiment analysis models, and the available models did not perform well on our data. Furthermore, there are no BERT models pretrained on clinical data for Dutch, so we used the multilingual BERT model as a basis. The positive sentiment model performs better than most other studies, with an F1 score of 0.97. Other studies report F1 scores between 0.74 and 0.90 for sentiment analysis on patient experience data[6,14,15,28,29]. The negative sentiment model performs below average, with an F1 score of 0.63. The small number of negative texts compared to the amount of neutral and positive texts causes this difference. With more data, the model can be trained further to improve the performance in recognizing negative texts and make it more generalizable to other departments and care pathways.

Our manual validation of the NLP pipeline shows that the quality of the topics is high in terms of the representativeness of the topics and the similarity to the manual topics. These results align with previous studies that show the similarity between supervised, manually defined topics and unsupervised, automatically defined topics[7,15]. However, there is a large difference in the quality of the topics for the different categories in the AI-PREM. Although most topics represent their texts very well with scores ranging from 90 to 100%, a few mostly negative topics have scores between 20 and 50%. One possible explanation is the heterogeneity in the negative answers, leading to a few 'left-over' topics that fail to represent the texts well. One solution would be to gather more data before running the model, as this would decrease the chance of getting topics that only contain a few texts. Another solution is changing the phrasing of the questionnaire by making it more specific or giving different examples. Especially the question about the organization could be improved because this question also showed low responsiveness to changes in sentiment. On the other hand, the number of texts that could not be assigned a topic was only 2.8%, which

is much better than the 15.4% reported in previous work[10]. It shows that a larger amount of texts can be automatically analyzed and confirms the improved suitability of our proposed open-ended questions for NLP analysis. In a previous report by Spasíc et al.[16], the authors optimize their questionnaire comprising open-ended questions in a similar way, i.e., by focusing every question on one particular aspect (different patient outcomes in their case), extracting any sentiment from the question itself, and providing examples per question (also at their patients' request).

We noted that positive comments are much more numerous, but negative topics tend to be more elaborately discussed by patients. For example, the negative topics' wait result scan' and 'contact (with) other hospital' contain concrete problems, while 'information good' and 'only positive' are much more high-level. These results align with other studies[3,11,30], which also found more specific feedback in negative comments. As we aimed to facilitate the quality improvement process, we see no limitation in this finding: the in-depth nature of the negative feedback makes it possible to define specific points of improvement, while the more general positive feedback functions as motivation for healthcare professionals. Moreover, previous work on structured patient experience questionnaires describes the problem of the ceiling effect: patient experience questionnaires tend to overestimate patient satisfaction[4], and very satisfied patients often still include a point of improvement[5,31]. The AI-PREM shows this same trend towards positive responses, but the ability to provide a free text response leads to more in-depth feedback. The tool further facilitates healthcare professionals to put topics into perspective by comparing positive to negative topics and forming concrete action points by going back to patients' original responses.

Strengths & limitations

A strength is the combination of quantitative data from the sentiment analysis and qualitative data from the topic models, which creates a clear, usable overview of patients' experiences. It also aligns with the proposed framework for automated analysis of opinionated data from a recent study[32]. This framework presents a similar pipeline, with sentiment analysis for the quantitative analysis followed by a more qualitative approach using, for example, topic modeling.

Another strength of the current study is the validation steps we took to assess the performance of the AI-PREM tool. Although it was challenging to find suitable validation methods, the current methods combined with the COSMIN reporting guideline provide some insight into how well the topics represent the patients' answers. However, the combination of the small sample size per topic and lack of easily interpretable metrics limits the use of topic modeling. Therefore, we could not compare our topic models to other literature.

The current sentiment analysis model, which assigns a whole text as either 'positive', 'neutral', or 'negative', is limited. By assigning texts as 'negative' if they contained at least one aspect that the patient was negative about, we made sure not to miss any points for improvement. However, in the future, we would like to finetune the model to define a sentiment per sentence instead of per text and to change the sentiment into a 5-point scale ranging from 'very dissatisfied' to 'very satisfied'. This granularity would make it easier to define priorities based on the level of dissatisfaction with a specific aspect of care.

Lastly, our current tool was built and validated in close consultation with clinicians, which ensures the internal validity of the model and clinically relevant and actionable output. However, it was validated using the patient experiences of a specific patient group. To investigate the generalizability of the AI-PREM tool, we will have to collect AI-PREM data in other patient groups and evaluate its usability for different groups of physicians.

Conclusions

The AI-PREM tool is a comprehensive method that combines a validated questionnaire consisting of open-ended questions with a well-performing NLP pipeline and visualization. By thematically organizing and quantifying patient feedback, it reduces the time invested by healthcare professionals to evaluate and prioritize patient experiences without being confined to the limited answer options of closed-ended questions.

Availability of data and materials

The datasets generated and analyzed during the current study are available from the corresponding author (MB) on reasonable request.

Author contributions

All authors contributed to the conceptualization of the project. E.F., O.N., and H.B. developed the questionnaire. M.B. developed the NLP model. M.B. and O.N. validated the questionnaire and NLP model. M.B. wrote the main manuscript text. All authors reviewed the manuscript.

Ethics declarations

The medical ethics committee from Leiden Den Haag Delft (METC LDD) waived the need for approval for this study. The reference number of the waiver is N19.121.

The authors declare that they have no competing interests.

Supplementary information

References

1. Bastemeijer CM, Boosman H, Zandbelt L, *et al.* Patient Experience Monitor (PEM): The Development of New Short-Form Picker Experience Questionnaires for Hospital Patients with a Wide Range of Literacy Levels. *Patient Relat Outcome Meas* 2020;**Volume 11**:221–30. doi:10.2147/prom.s274015
2. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Published Online First: 2001. doi:10.17226/10027
3. Khanbhai M, Anyadi P, Symons J, *et al.* Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *Bmj Heal Care Informatics* 2021;**28**:e100262. doi:10.1136/bmjhci-2020-100262
4. Riiskjaer E, Ammentorp J, Kofoed P-E. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. *Int J Qual Health C* 2012;**24**:509–16. doi:10.1093/intqhc/mzs039
5. Alemi F, Torii M, Clementz L, *et al.* Feasibility of Real-Time Satisfaction Surveys Through Automated Analysis of Patients' Unstructured Comments and Sentiments. *Qual Manag Health Ca* 2012;**21**:9–19. doi:10.1097/qmh.0b013e3182417fc4
6. Anjum A, Zhao X, Bahja M, *et al.* Identifying patient experience from online resources via sentiment analysis and topic modelling. *Proc 3rd Ieee Acm Int Conf Big Data Comput Appl Technologies* 2016;:94–9. doi:10.1145/3006299.3006335
7. Jones J, Pradhan M, Hosseini M, *et al.* Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. *Jmir Medical Informatics* 2018;**6**:e45. doi:10.2196/medinform.9162
8. Greaves F, Ramirez-Cano D, Millett C, *et al.* Machine learning and sentiment analysis of unstructured free-text information about patient experience online. *Lancet* 2012;**380**:S10. doi:10.1016/s0140-6736(13)60366-9
9. Ranard BL, Werner RM, Antanavicius T, *et al.* Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. *Health Affair* 2017;**35**:697–705. doi:10.1377/hlthaff.2015.1030
10. Cammel SA, Vos MSD, Soest D van, *et al.* How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *Bmc Med Inform Decis* 2020;**20**:97. doi:10.1186/s12911-020-1104-5
11. Khanbhai M, Warren L, Symons J, *et al.* Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. *Int J Med Inform* 2022;**157**:104642. doi:10.1016/j.ijmedinf.2021.104642
12. Menendez ME, Shaker J, Lawler SM, *et al.* Negative Patient-Experience Comments After Total Shoulder Arthroplasty. *J Bone Joint Surg* 2019;**101**:330–7. doi:10.2106/jbjs.18.00695
13. Rivas C, Tkacz D, Antao L, *et al.* Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. *Heal Serv Deliv Res* 2019;**7**:1–160. doi:10.3310/hsdr07230

14. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. *Appl Clin Inform* 2020;**11**:242–52. doi:10.1055/s-0040-1708049

15. Doing-Harris K, Mowery DL, Daniels C, *et al.* Understanding patient satisfaction with received healthcare services: A natural language processing approach. In: *AMIA Annual Symposium Proceedings*. 2017. doi:28269848

16. Spasić I, Owen D, Smith A, *et al.* KLOSURE: Closing in on open–ended patient questionnaires with text mining. *J Biomed Semant* 2019;**10**:24. doi:10.1186/s13326-019-0215-3

17. Davis K, Schoenbaum SC, Audet A-M. A 2020 vision of patient-centered primary care. *J Gen Intern Med* 2005;**20**:953–7. doi:10.1111/j.1525-1497.2005.0178.x

18. Soulier G, Leeuwen BM van, Putter H, *et al.* Quality of Life in 807 Patients with Vestibular Schwannoma: Comparing Treatment Modalities. *Otolaryngology Head Neck Surg* 2017;**157**:92–8. doi:10.1177/0194599817695800

19. Gagnier JJ, Lai J, Mokkink LB, *et al.* COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res* 2021;**30**:2197–218. doi:10.1007/s11136-021-02822-4

20. Hugging Face. BERT. https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification (accessed 14 Dec 2021).

21. Norvig P. How to Write a Spelling Corrector. 2016.https://norvig.com/spell-correct.html (accessed 21 Nov 2021).

22. Seal M, Rodriguez T. *CyHunSpell*. 2021. https://pypi.org/project/cyhunspell/

23. Qi P, Zhang Y, Zhang Y, *et al.* Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

24. Tulkens S, Emmery C, Daelemans W. Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC 2016*). European Language Resources Association (ELRA) 2016.

25. Schäfer R, Bildhauer F. Building Large Corpora from the Web Using a New Efficient Tool Chain. 23AD;:486–93.http://rolandschaefer.net/?p=70

26. Schäfer R. Processing and querying large web corpora with the COW14 architecture. Published Online First: 2015.http://rolandschaefer.net/?p=749

27. Council of Europe. Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume. Strasbourg: : Council of Europe Publishing 2020. www.coe.int/lang-cefr

28. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics* 2020;**8**. doi:10.2196/17984

29. Jiménez-Zafra SM, Martín-Valdivia MT, Maks I, *et al.* Analysis of patient satisfaction in Dutch and Spanish online reviews. *Procesamiento del Lenguaje Natural* 2017;:101–8.

30. Wagland R, Recio-Saucedo A, Simon M, *et al.* Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. *Bmj Qual Saf* 2016;**25**:604. doi:10.1136/bmjqs-2015-004063

31. Gallan AS, Girju M, Girju R. Perfect ratings with negative comments: Learning from contradictory patient survey responses. *Patient Exp J* 2017;**4**:15–28. doi:10.35680/2372-0247.1234

32. Kazmaier J, Vuuren JH van. A generic framework for sentiment analysis: Leveraging opinion-bearing data to inform decision making. *Decis Support Syst* 2020;**135**:113304. doi:10.1016/j.dss.2020.113304

Chapter 5

Applying Natural Language Processing to Patient Messages to Identify Depression Concerns in Cancer Patients

Marieke M. van Buchem, Anne A.H. de Hond, Claudio Fanconi, Vaibhavi Shah, Max Schuessler, Ilse M.J. Kant, Ewout W. Steyerberg, Tina Hernandez-Boussard

Journal of the American Medical Informatics Association, 17 July 2024

5.1 Abstract

Objective

This study aims to explore and develop tools for early identification of depression concerns among cancer patients by leveraging the novel data source of messages sent through a secure patient portal.

Materials and Methods

We developed classifiers based on logistic regression (LR), support vector machines (SVMs), and 2 Bidirectional Encoder Representations from Transformers (BERT) models (original and Reddit-pretrained) on 6600 patient messages from a cancer center (2009-2022), annotated by a panel of healthcare professionals. Performance was compared using AUROC scores, and model fairness and explainability were examined. We also examined correlations between model predictions and depression diagnosis and treatment.

Results

BERT and RedditBERT attained AUROC scores of 0.88 and 0.86, respectively, compared to 0.79 for LR and 0.83 for SVM. BERT showed bigger differences in performance across sex, race, and ethnicity than RedditBERT. Patients who sent messages classified as concerning had a higher chance of receiving a depression diagnosis, a prescription for antidepressants, or a referral to the psycho-oncologist. Explanations from BERT and RedditBERT differed, with no clear preference from annotators.

Discussion

We show the potential of BERT and RedditBERT in identifying depression concerns in messages from cancer patients. Performance disparities across demographic groups highlight the need for careful consideration of potential biases. Further research is needed to address biases, evaluate real-world impacts, and ensure responsible integration into clinical settings.

Conclusion

This work represents a significant methodological advancement in the early identification of depression concerns among cancer patients. Our work contributes to a route to reduce clinical burden while enhancing overall patient care, leveraging BERT-based models.

5.2 Background and significance

Depression is common in cancer patients and negatively associated with treatment outcomes, prognosis, and quality of life[1–5]. Despite its prevalence in cancer patients (20% in the United States[6]), depression often remains underdiagnosed. This leads to delayed intervention, poorer treatment adherence, and potential exacerbation of the patient's overall health status[1–3,7–9]. Early identification of depression symptoms may facilitate timely mental health support.

A majority of tools for depression screening in cancer patients utilize structured surveys. Many of these tools perform well in the identification of cancer patients with depression. However, most clinicians do not use structured depression scales during routine clinical care, as they perceive them as too long[10]. To address this, machine learning (ML) approaches using clinical data have also been explored[11]. For example, Cho et al. trained an ML model to predict depression using nationwide clinical check-up data, attaining an AUC of 0.84[12]. While some ML tools demonstrate promising performance, they primarily rely on clinician-generated data, which may not fully capture the patient's perspective or experiences. This limitation highlights the need for alternative data sources that can provide a more comprehensive view of the patient's mental health status.

Patient-generated health data, such as messages sent through secure patient portals, present a unique yet underutilized source of information for identifying signs of depression. These messages, often exchanged between patients and healthcare providers, can provide insights into the patient's mental health status, potentially enabling early detection and treatment of depression symp-

toms. However, the increasingly high number of patient messages also leads to challenges for healthcare professionals. Previous studies suggest that the high volume of clinical communications can lead to professional exhaustion and burnout[13–15]. Applying natural language processing (NLP) to patient-generated health data might offer a solution by monitoring all incoming messages for potential signs of depression.

Most previous work on applying NLP to identify mental health issues in patient-generated health data is focused on social media data[16–25]. Social media is a valuable source, as it provides extensive documentation of people's personal thoughts, experiences, and ideas. Especially Reddit is an interesting medium as it is fully anonymous, enabling honest conversations between users[26,27]. However, the downside of detecting mental health issues through these kinds of media is that it is difficult to provide support to the individual users. A recent study described the development of an NLP model to predict suicide-related events from patient portal messages, showing promising results comparable to commonly used assessment tools[28]. Therefore, it might be possible to assist healthcare professionals in identifying cancer patients that potentially suffer from depression.

Objective

This study aims to develop a proof-of-concept model using NLP to analyze incoming patient messages and identify those messages indicative of depression concerns. We compare classical machine learning approaches with neural networks-based Bidirectional Encoder Representations from Transformers (BERT) models[29] and investigate whether domain-adaptive pretraining on Reddit data improves the performance of the model.

5.3 Methods

Data

The dataset consisted of patient-initiated messages that were sent through a secure patient portal by patients from a comprehensive cancer cohort, contain-

ing all patients who visited the Stanford Cancer Center from 2009 until 2022. The secure patient portal allows patients to send messages to their care team. We only included English, patient-initiated messaging threads and excluded all standard communication, e.g. reminders for appointments and invitations for patient satisfaction questionnaires. We did not exclude patients with a prior depression diagnosis, as this has previously been found to be the most predictive factor for developing depression[30,31]. We aimed for a final sample size of at least 5000 manually annotated messages, based on similar studies leveraging BERT for binary text classification on manually annotated data[32–34]. For the final annotation sample, we randomly sampled 50% of messages from the dataset. To decrease the imbalance in the sample towards non-concerning messages, we selected the other 50% of the annotation sample to contain potentially alarming or concerning content. To this end, an experienced social worker and data scientist created two lists of words that signaled concern or alarm respectively (see Appendix A). The selected sample included all the messages containing alarming words, supplemented with randomly sampled messages containing concerning words. This distribution was chosen to maximize the number of potentially concerning messages within the annotation sample. Additionally, special attention was given to discerning depression from anxiety. Our word lists included terms related to both depression and anxiety to ensure these messages were manually annotated, enhancing the model's capability to distinguish messages expressing anxiety from potential depression.

Ethical Considerations

This study was approved by the Stanford institutional review board (#47644). Informed consent was waived for this retrospective study for access to personally identifiable health information as it would not be reasonable, feasible, or practical. The data are housed in the Stanford Nero Computing Platform, which is a highly secure, fully integrated internal research data platform meeting all security standards for high risk and protected health information data. The security is managed and monitored, and the platform is updated and adapted to meet regulatory changes.

Annotation process

The definition of when a message was 'concerning for depression' was created through multi-stakeholder input (see Appendix B). We organized brainstorm sessions with oncologists, data scientists, medical students, and a social worker, during which we iteratively worked towards a definition and annotation guideline that everyone agreed on (see Appendix B). When consensus was reached, the annotation was performed by seven healthcare professionals. We created a diverse group of annotators, in terms of clinical experience, age, gender, race, and ethnicity. The final set of 6,600 annotated messages was used as the reference standard. Of 6,600 messages, a set of 100 messages was annotated by all annotators. We computed the inter annotator agreement on this sample using Krippendorf's Alpha. We used majority vote to define the reference standard for these 100 messages. See Figure 1 for an overview of the annotation process.

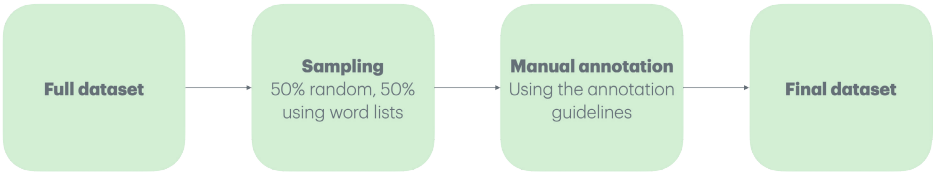


Figure 1: overview of the sampling and annotation pipeline.

Models

For this study, we aimed to compare the performance of two baseline machine learning models (logistic regression [LR] and support vector machines [SVM]) to two bidirectional encoder representations from transformers (BERT) models: a naïve BERT base model and a BERT base model that has undergone continued domain-adaptive pretraining on Reddit data. The method of continued domain-adaptive pretraining is computationally less expensive than a full pretraining task, while it has shown to improve the performance on specific tasks[35–40]. We chose the BERT base in combination with continued pretraining on Reddit data, as opposed to ClinicalBERT[41] or BioBERT[36], because Reddit data includes a wide range of discussions, including those related to personal experiences and mental health, which are closer to the type of

conversational and informal language found in patient portal messages. Patient portal messages often reflect patients' everyday language and concerns, which may not be captured in more formal medical records or biomedical literature. This similarity in language and context can help the model better understand and interpret patient messages. We specifically used the Depression subreddit to include the type of language that people use to talk about depression.

For the LR and SVM models, we first preprocessed the data by changing all letters to lowercase, removing stop words, and stemming the words. We used the term frequency – inverse document frequency (TF-IDF) to extract features from the data. The final dataset was split into 70-15-15 train-validation-test sets. To find the best hyperparameters for the TF-IDF vectorizer, the LR and the SVM, we performed a grid search using the train and validation set. We then fit the LR and the SVM model with the best hyperparameters and performed a bootstrap with 1000 samples using the test set to determine the performance of the models.

For our domain-adaptive pretraining task, we chose a specific Reddit community ('subreddit') called 'r/Depression'. This is a large subreddit with more than 900,000 members that has been in use since 2009 and, therefore, provides extensive data on a large time span. It is the biggest subreddit focused on depression, with millions of posts. From this subreddit, we scraped 1,000,000 posts. We then continued pretraining the BERT base model for 20 epochs [35]. The pretraining was performed using four GPUs on the Google Cloud Platform. The final model is referred to as RedditBERT. Both BERT base and RedditBERT were finetuned on the binary classification task of identifying concerning messages, using the annotated sample of patient messages.

Associations between model predictions and patient characteristics

We conducted a comparative analysis of patient characteristics and clinical outcomes between patients who sent messages deemed concerning by RedditBERT, and those who did not. Included outcomes were a depression diagnosis, prescriptions for depression medication, and mental health referrals (Appendix C). Differences in categorical variables were assessed using a chi-square test. For continuous variables, an independent samples T-Test was performed. In cases

where continuous variables encompassed more than two groups, a one-way Analysis of Variance (ANOVA) was performed.

Explainability

We used Local Interpretable Model-Agnostic Explanations (LIME) to compare BERT and RedditBERT explanations on a patient message level[42]. LIME provides the local importance of each word to the model's classification of a specific patient message. Our sample included 32 texts categorized into four distinct buckets based on the outputs of the two BERT models (BERT and RedditBERT) and their alignment with the reference standard (human annotation) (see Figure 2). Subsequently, our seven annotators were asked to compare the predictions generated by the two models and the corresponding LIME explanations. Through a structured survey, annotators were asked to indicate which prediction they agreed with and which explanation they preferred (Appendix D). Additionally, the survey provided an opportunity for the annotators to explain their decisions.

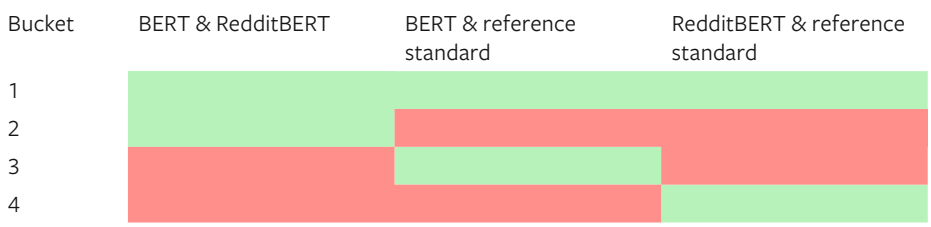


Figure 2: description of the four buckets used for evaluation of the explainability. An empty cell indicates agreement, a diagonal line indicates disagreement.

5.4 Results

Patient characteristics

The total data set included 6,600 messages from 3,312 unique patients. The cohort consisted of more females (60%), an average age of 61 years old and a majority of White and Asian, privately insured patients (see Table 1 for more characteristics). Our final test set consisted of 907 messages (14% of the total

labeled set) from 760 unique patients. 90 messages (10%) were annotated as concerning for depression.

Table 1. Patient characteristics cohort.

	N=3,312
Demographics	
Female sex, N(%)	2,002 (60)
Age, mean (std)	61.3 (13.7)
English speaking, N(%)	3,080 (93)
Race	
Asian (%)	725 (22)
Black (%)	65 (2)
White (%)	2,133 (64)
Other (%)	364 (11)
Ethnicity	
Hispanic/Latino (%)	204 (6)
Non-hispanic/non-latino (%)	3,060 (92)
Other (%)	47 (1)
Depression diagnosis, N(%)	1,116 (34)
Insurance (%)	
Private	1,980 (60)
Medicare	550 (17)
Medicaid	260 (8)
Other	522 (16)

Inter-annotator agreement

The inter annotator agreement (IAA) calculated over all seven annotators was 0.38 according to Krippendorf's Alpha, which can be considered moderate. We observed a large variation in IAA between different sets of annotators, ranging from 0.32 to 0.52, depending on which annotator was removed from the set.

Table 2. Performance metrics of four models classifying patient messages as concerning for depression.

Metric, mean [95% CI] based on 1000 bootstraps	Log Reg *Threshold: 0.2**	SVM *Threshold: 0.5**	BERT *Threshold: 0.5**	RedditBERT *Threshold: 0.5**
AUROC	0.79 [0.74-0.83]	0.83 [0.78-0.87]	0.86 [0.82-0.90]	0.88 [0.85-0.91]
Precision	0.32 [0.25-0.39]	0.36 [0.28-0.44]	0.37 [0.30-0.44]	0.33 [0.26-0.39]
Recall	0.51 [0.40-0.61]	0.60 [0.49-0.70]	0.68 [0.59-0.78]	0.74 [0.66-0.84]
F1-score	0.39 [0.31-0.47]	0.45 [0.37-0.52]	0.48 [0.40-0.55]	0.46 [0.39-0.53]

* = *threshold chosen that led to the highest F1 score*

Table 3. Predictive performance per subgroup.

	BERT AUC [95% CI]	Recall [95% CI]	RedditBERT AUC [95% CI]	Recall [95% CI]
Overall	0.86 [0.82-0.90]	0.69 [0.59-0.78]	0.88 [0.85-0.91]	0.74 [0.65-0.83]
Sex				
Female (n=476)	0.85 [0.79-0.91]	0.73 [0.61-0.84]	0.88 [0.84-0.92]	0.73 [0.61-0.84]
Male (n=284) 0.89 [0.83-0.94] 0.62 [0.43-0.78] 0.90 [0.84-0.94] 0.76 [0.62-0.90]				
Race				
Asian (n=136)	0.87 [0.74-0.98]	0.75 [0.53-0.95]	0.91 [0.83-0.97]	0.63 [0.37-0.86]
Black (n=16)	0.82 [N/A]	0.33 [N/A]	0.75 [N/A]	0.33 [N/A]
White (n=519)	0.86 [0.81-0.90]	0.67 [0.54-0.79]	0.88 [0.85-0.92]	0.79 [0.68-0.89]
Other (n=81)	0.83 [0.64-0.98]	0.74 [0.44-1.00]	0.90 [0.79-0.98]	0.75 [0.44-1.00]
Ethnicity				
Hispanic/Latino (n=44)	0.80 [0.54-0.99]	0.66 [0.33-1.00]	0.88 [0.73-0.99]	0.78 [0.44-1.00]
Non-Hispanic/non-Latino (n=709)	0.87 [0.83-0.91]	0.69 [0.58-0.79]	0.89 [0.86-0.92]	0.75 [0.66-0.84]
Other (n=7)	0.95 [N/A]	0.67 [N/A]	0.95 [N/A]	0.33 [N/A]

Model performance

The TF-IDF parameters and hyperparameters of the logistic regression (LR) and support vector machine (SVM) can be found in Appendix E. The LR model had a mean area under the ROC curve (AUROC) of 0.79 (95% confidence interval (CI): 0.74-0.83) while the SVM attained an AUROC of 0.83 (95% CI: 0.78-0.87).

Both BERT and RedditBERT were trained and validated for 5 epochs on 5,693 labeled messages. See Appendix E for hyperparameters. Both models outperformed the LR and SVM and RedditBERT slightly outperformed BERT, with a mean AUROC of 0.88 (95% CI: 0.85-0.91) versus 0.86 (95% CI: 0.82-0.90), respectively. A threshold of 0.5 led to the highest F1 score for the BERT models and the SVM. For the LR, a threshold of 0.2 led to the highest F1 score (Table 2). In total, RedditBERT labeled 200 messages as concerning (22%). When comparing the predictive performance per subgroup, BERT showed bigger differences in performance across sex, race, and ethnicity than RedditBERT. For both models there was a decreased performance for Black patients (Table 3).

Associations between model predictions and patient characteristics

There was a significant difference in race in the classification of patients' messages. Messages of White patients were more often classified as concerning, while messages of Asian patients were less often classified as concerning (see Appendix F). Furthermore, patients on Medicaid or Medicare also sent more messages classified as concerning. Patients who sent messages classified as concerning by RedditBERT had a higher chance of receiving a depression diagnosis, a prescription for antidepressants, or a mental health referral within the next 3, 6, and 12 months after sending the concerning message. Patients sending a concerning message were also more likely to already have a depression diagnosis, a prescription for antidepressants, or a mental health referral (see Appendix F).

Explainability

The explanation of which words contributed to the prediction per message differed for BERT and RedditBERT, with RedditBERT highlighting more words than BERT (see Appendix G). Annotators preferred BERT's explanation to RedditBERT's explanations for 14 out of 26 texts (54%). Annotators often opted for Red-

ditBERT's explanation when it highlighted words or sentences that BERT missed. On the other hand, annotators sometimes preferred BERT's explanation because they found RedditBERT highlighted words that did not make sense in the eyes of the annotators. Furthermore, several annotators mentioned that the words highlighted as 'not concerning' did not always seem to make sense to them (Table 4).

Table 4. Annotators' reasons for choosing BERT or RedditBERT's explanation.

Reasons for choosing RedditBERT	Reasons for choosing BERT
"Difficult. I like the explanations of [RedditBERT] a bit more, because it seems to pick out more complete sentences like ' am extremely tired' and 'have not ... able to sleep more'."	"I prefer [BERT] because the blue [non-concerning] words in [RedditBERT], do not make sense to me. Why should testosterone be marked as non-concerning."
"This is the best use case of this model. A clear cry for help. I like the explanations of [RedditBERT] better because it picks out more complete sentences 'I'm pretty depressed' and has a stronger reaction on the 'psychologist'."	"[RedditBERT] highlights a lot of text that I do not think relevant in either direction."
"I think in general it is good [RedditBERT] picks up on prescription names."	"Again, there is a lot of text highlighted in both that does not really make sense to me. [BERT] highlights less text."
"I like how [RedditBERT] picks up on the 'love to talk to somebody'. "	"I do not agree with the extra highlighted words really in [RedditBERT], as the only indication of concern is the 'depressed'."

5.5 Discussion

In this study, we demonstrate a proof-of-concept for leveraging patient-generated health data for the early identification of depression concerns in cancer patients. By employing natural language processing (NLP) techniques, specifically Bidirectional Encoder Representations from Transformers (BERT) and domain-adaptive pretraining using Reddit data (RedditBERT), we highlight the potential of artificial intelligence in enhancing mental health surveillance. However, the performance disparities observed across patient subgroups, notably concerning race and ethnicity, necessitate a careful consideration of the ethical implications and potential biases introduced by these models.

The good discriminatory ability across all models showed the potential of patient messages as a valuable source for depression risk stratification. Our results are comparable to one other study that used patient portal messages to identify a mental health event, namely suicide [28]. For this study, the authors reported an AUROC of 0.71. Both findings underline the potential of using patient messages as a unique data source, which provides a current snapshot of how a patient is feeling and directly represents the patient's voice. This untapped data source has the opportunity to improve personalized, proactive identification of mental health issues. However, more research on this topic is needed, as these are the only studies describing the application of NLP on this data source.

There was no significant difference between the naïve base BERT and the domain-adaptive pretrained RedditBERT model. This finding contradicts previous studies in which domain-specific models like BioBERT (pre-trained on biomedical texts) and ClinicalBERT (pretrained on clinical texts), and continuously pretrained BERT models outperformed base BERT [35,36,41,43]. However, within the mental health domain, a recent study also found that depression classification did not improve significantly with continued pretraining[44]. More research is needed to assess the value of social media data for continued pretraining in the mental health domain.

We found a notable difference in how words were weighed in the explanations provided for BERT and RedditBERT, but there was no difference on average in the quality of the explanations. Explanations may help generate trust in deep neural network models, like BERT, which are inherently uninterpretable[45]. Yet, post-hoc explainability methods like LIME are difficult to validate, and their effect on clinical decision making is still unknown. More research is needed on the added value of explainability methods in increasing trust versus the potential to harm trust[46]. Alternatively, neural networks with a more inherent interpretability mechanism could lead to better explanations[47].

The subgroup analysis showed slight differences in performance between sex, race, and ethnicity. Compared to BERT, RedditBERT performed more consistently between subgroups and had a slightly better recall for male patients

and White patients, which could be due to Reddit being predominantly used by males[48]. Furthermore, both models performed worse on Black patients, which can be explained by the low number of Black patients within our sample. This finding highlights the importance of addressing potential biases and ethical considerations associated with deploying AI models in healthcare, emphasizing the need for equitable and unbiased implementations. The National Institute of Health's All of Us Research Program is a great example of an initiative aiming to collect data from a diverse group of participants across the US[49]. For future research, we recommend training models on such a diverse dataset to decrease differences in subgroup performance.

A depression diagnosis or prescription of depression medication occurred more often after a concerning message was sent compared to after a non-concerning message was sent. This suggests that our models were able to identify messages that were truly indicative of depression concerns. These may be patients that could benefit from additional mental healthcare outreach. Important to note, however, is that some of these patients already received a depression diagnosis or treatment. This highlights the classification capabilities of the model, although the model might not perform well for prediction. This assumption is underlined by a recent study, where we show that using the output of our model does not improve the performance of a prediction model for depression[31].

Limitations

One limitation is the moderate inter-annotator agreement. This can be attributed to the diversity among the annotators and the inherent subjective interpretation of what qualifies as a 'concerning' message in patient emails. This is highlighted by the large variation in IAA, depending on which annotators are included. Although we believe the IAA could be improved by excluding some annotators, it also mirrors the real-world where different healthcare professionals may interpret patient communications differently. Relevant literature describing similar use cases, such as annotating Twitter data for mental health symptoms, report similar moderate inter-annotator agreements[50,51]. Despite the moderate agreement, the study's rigorous approach in involving multiple annotators and the alignment with existing literature provide valuable insights

into the complexities of labeling subjective content. Taking this into account, we conclude that the use of patient messages combined with labels from our diverse, clinical group of annotators greatly improved the method's potential to be applicable in healthcare practice.

Furthermore, our current approach to upsampling concerning messages may lead to a bias in the training data towards messages that are more easily identifiable as concerning. As the current study is a proof-of-concept, we chose this method to keep the manual annotation feasible while still ensuring that there were enough concerning messages to train the model effectively. However, future work should explore more sophisticated sampling techniques to better represent the full spectrum of patient messages and minimize potential biases.

Another limitation of this study is the focus on a single institution, which may limit the generalizability of our results to other settings. Especially the high number of privately insured patients is not representative for the general population. Nevertheless, this cohort included a diverse population in terms of race and ethnicity, with a substantial percentage of Hispanic and Asian patients. This study can thus be seen as a proof-of-concept and sets the stage for future investigations into the ways different ethnic and cultural groups express mental health concerns in their communications. By recognizing and addressing these differences, subsequent studies can delve deeper into tailoring interventions that resonate effectively across diverse patient populations.

Lastly, the study's framework might not capture patients who do not use emails for communication or are hesitant to reach out, thereby potentially missing a subset of the population in need.

Future implications

Given our exploration in the use of advanced NLP models for the identification of depression concerns in cancer patients, the broader implications for healthcare are significant. The advantage of BERT and RedditBERT over traditional methods underscores the potential of integrating more sophisticated language models into clinical practice. With the ongoing advancements in NLP, especially

in the field of large language models (LLM's), there is the potential to further refine these models, making them even more relevant and effective in a clinical context. Future work should focus on comparing several newer language models to determine if they could provide improved performance in identifying depression. Recent studies have shown that it is also possible to use LLMs to create chatbots for counseling, offering another promising avenue for providing mental health support[52]. However, while the promise of these advanced NLP models in healthcare is evident, it's crucial to approach their integration with caution. Before such models can be responsibly incorporated into clinical settings, additional research is required to address potential biases as were demonstrated in the current study and evaluate the real-world impact on physician-patient interaction and clinical outcomes[53–55].

Furthermore, our study significantly contributes to the literature by emphasizing the underutilized potential of patient-generated health data, specifically messages sent through a secure patient portal. This novel approach taps into valuable information exchanged between patients and healthcare providers, offering insights into the mental health state of a patient and enabling early detection of depression concerns. This data is systematically collected, as opposed to, for example, patient reported outcomes (PROs). The collection of PROs is often burdensome to patients and healthcare providers, may not capture all patients' concerns, and rely on patients' memory to report symptoms that have occurred prior to the patient's visit.18 Thus, patient messages should be seen as a valuable additional data source for clinical research and surveillance.

Following our proof-of-concept study, we propose several next steps. First of all, to ensure broader applicability of such a tool, the training dataset should be extended with data representative of the general population. Secondly, it is important to conduct a temporal validation to assess the model's performance over time. Lastly, other types of explainability methods should be tested to determine if some provide a better understanding of the model's behavior than the current method. These steps will help refine the model further and enhance its applicability and trustworthiness in a clinical setting.

Conclusion

In conclusion, this work represents a significant methodological advancement in the early identification of depression concerns among cancer patients, addressing a critical gap in patient care. Our work contributes to a route to reduce clinical burden while enhancing overall patient care, leveraging BERT-based models. Further research is needed to address biases, evaluate real-world impacts, and ensure responsible integration into clinical settings. As the study highlights, the interpretability of these models is paramount for clinician trust and responsible implementation in healthcare settings, particularly for vulnerable patient populations.

Author contributions

Marieke M. van Buchem, Anne A.H. de Hond, Ewout W. Steyerberg, Ilse M.J. Kant, and Tina Hernandez-Boussard were responsible for the conceptualization and design of the study. Marieke M. van Buchem and Anne A.H. de Hond performed the data extraction. Marieke M. van Buchem performed the data analysis. Max Schuessler and Vaibhavi Shah provided clinical advice. Marieke M. van Buchem drafted the original manuscript. All authors had full access to all the data, critically analyzed, reviewed, contributed, and approved the final manuscript.

Funding

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR003142 and the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013362. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.v.B. and A.d.H. received a travel grant from the Cath-

arine van Tussenbroek Fund and the Prins Bernhard Cultuur Fund to support this research.

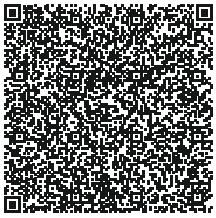
Conflicts of interest

The authors have no competing interests to disclose.

Data availability

The data underlying this article cannot be shared due to the sensitivity of the content and the privacy of individuals that participated in the study.

Supplementary information



References

1. Linden W, Vodermaier A, MacKenzie R, *et al.* Anxiety and depression after cancer diagnosis: Prevalence rates by cancer type, gender, and age. *J Affect Disord*. 2012;141:343–51.
2. SMITH HR. Depression in cancer patients: Pathogenesis, implications and treatment (Review). *Oncol Lett*. 2015;9:1509–14.
3. Pitman A, Suleman S, Hyde N, *et al.* Depression and anxiety in patients with cancer. *BMJ*. 2018;361:k1415.
4. Colleoni M, Mandala M, Peruzzotti G, *et al.* Depression and degree of acceptance of adjuvant cytotoxic drugs. *Lancet*. 2000;356:1326–7.
5. Grassi L, Indelli M, Marzola M, *et al.* Depressive symptoms and quality of life in home-care-assisted cancer patients. *J Pain Symptom Manag*. 1996;12:300–7.
6. HHS SA and MHSA (SAMHSA). Substance Abuse and Mental Health Services Administration; mental health and substance abuse emergency response criteria. Interim final rule. *Fed Regist*. 2001;66:51873–80.
7. Walker J, Hansen CH, Martin P, *et al.* Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. *Lancet Psychiatry*. 2014;1:343–50.
8. Caruso R, Breitbart W. Mental health care in oncology. Contemporary perspective on the psychosocial burden of cancer and evidence-based interventions. *Epidemiology Psychiatr Sci*. 2020;29:e86.
9. Mitchell AJ, Chan M, Bhatti H, *et al.* Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. *Lancet Oncol*. 2011;12:160–74.
10. Mitchell AJ, Meader N, Davies E, *et al.* Meta-analysis of screening and case finding tools for depression in cancer: Evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *J Affect Disord*. 2012;140:149–60.
11. Iyortsuun NK, Kim S-H, Jhon M, *et al.* A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare*. 2023;11:285.
12. Cho S-E, Geem ZW, Na K-S. Prediction of depression among medical check-ups of 433,190 patients: A nationwide population-based study. *Psychiatry Res*. 2020;293:113474.
13. Tai-Seale M, Dillon EC, Yang Y, *et al.* Physicians' Well-Being Linked To In-Basket Messages Generated By Algorithms In Electronic Health Records. *Heal Aff*. 2019;38:1073–8.
14. Adler-Milstein J, Zhao W, Willard-Grace R, *et al.* Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med Inform Assoc*. 2020;27:531–8.
15. Lieu TA, Altschuler A, Weiner JZ, *et al.* Primary Care Physicians' Experiences With and Strategies for Managing Electronic Messages. *JAMA Netw Open*. 2019;2:e1918287.
16. Arachchige IAN, Sandanapitchai P, Weerasinghe R. Investigating Machine Learning & Natural Language Processing Techniques Applied for Predicting Depression Disorder from Online Support Forums: A Systematic Literature Review. *Information*. 2021;12:444.

17. Tejaswini V, Babu KS, Sahoo B. Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model. *ACM Trans Asian Low-Resour Lang Inf Process*. 2022;23(1):1-20.
18. Katchapakirin K, Wongpatikaseree K, Yomaboot P, *et al.* Facebook Social Media for Depression Detection in the Thai Community. *2018 15th Int Jt Conf Comput Sci Softw Eng (JCSSE)*. IEEE; 2018;00:1–6.
19. Asad NA, Pranto MdAM, Afreen S, *et al.* Depression Detection by Analyzing Social Media Posts of User. *2019 IEEE Int Conf Signal Process, Inf, Commun Syst (SPICSCON)*. IEEE; 2019;00:13–7.
20. Kabir MK, Islam M, Kabir ANB, *et al.* Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques. *JMIR Form Res*. 2022;6:e36118.
21. Dessai S, Usgaonkar SS. Depression Detection on Social Media Using Text Mining. *2022 3rd Int Conf Emerg Technol (INCET)*. IEEE; 2022;00:1–4.
22. Haque A, Reddi V, Giallanza T. Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction. In*: Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks*. Springer International Publishing; 2021:436-447.
23. Ren L, Lin H, Xu B, *et al.* Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation. *JMIR Med Inform*. 2021;9:e28754.
24. Podina IR, Bucur A-M, Todea D, *et al.* Mental health at different stages of cancer survival: a natural language processing study of Reddit posts. *Front Psychol*. 2023;14:1150227.
25. Chen Z, Yang R, Fu S, *et al.* Detecting Reddit Users with Depression Using a Hybrid Neural Network. In*: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE; 2023:193-199.
26. Choudhury MD, De S. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *ICWSM*. 2014;8:71–80.
27. Ammari T, Schoenebeck S, Romero D. Self-declared Throwaway Accounts on Reddit: How Platform Affordances and Shared Norms enable Parenting Disclosure and Support. *Proc ACM Hum-Comput Interact*. 2019;3:1–30.
28. Bhandarkar AR, Arya N, Lin KK, *et al.* Building a Natural Language Processing Artificial Intelligence to Predict Suicide-Related Events Based on Patient Portal Message Data. *Mayo Clin Proc: Digit Heal*. 2023;1:510–8.
29. Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics; 2019:4171-4186.
30. Riedl D, Schüßler G. Factors associated with and risk factors for depression in cancer patients – A systematic literature review. *Transl Oncol*. 2022;16:101328.
31. Hond A de, Buchem M van, Fanconi C, *et al.* Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study. *JMIR Med Inform*. 2024;12:e51925.

32. Sousa MG de, Sakiyama K, Rodrigues L de S, *et al.* BERT for Stock Market Sentiment Analysis. In*: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE; 2019:1597-1601.
33. Du J, Xiang Y, Sankaranarayanapillai M, *et al.* Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. *J Am Med Inform Assoc*. 2021;28:1393–400.
34. Zhou S, Wang N, Wang L, *et al.* CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J Am Med Inform Assoc*. 2022;29:1208–16.
35. Lamproudis A, Henriksson A, Dalianis H. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In*: Proceedings of the International Conference on Recent Advances in Natural Language Processing—Deep Learning for Natural Language Processing Methods and Application*. INCOMA, Ltd.; 2021:790-797.
36. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
37. Gururangan S, Marasović A, Swayamdipta S, *et al.* Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In*: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:8342-8360.
38. Alsentzer E, Murphy JR, Boag W, *et al.* Publicly Available Clinical BERT Embeddings. In*: Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78.
39. Chakrabarty T, Hidey C, McKeown K. IMHO Fine-Tuning Improves Claim Detection. In*: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*). Association for Computational Linguistics; 2019:558-563.
40. Fanconi C, Buchem M van, Hernandez-Boussard T. Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes. *AMIA Jt Summits Transl Sci Proc*. 2023:138-147.
41. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Arxiv. 2020. Accessed July 12, 2024. https://arxiv.org/abs/1904.05342.
42. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In*: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics; 2016:97-101.
43. Peng B, Chersoni E, Hsu Y-Y, *et al.* Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. In*: Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics; 2021:37-44.
44. Ji S, Zhang T, Ansari L, *et al.* MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In*: Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association; 2022:7184-7190.
45. Amann J, Vetter D, Blomberg SN, *et al.* To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Heal*. 2022;1:e0000016.

46. Wysocki O, Davies JK, Vigo M, *et al.* Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif Intell*. 2023;316:103839.

47. Fanconi C, Vandenhirtz M, Husmann S, *et al.* This Reads Like That: Deep Learning for Interpretable Natural Language Processing. In*: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:14067-14076.

48. Reddit.com. Advertising—Audience—Reddit. Discover what makes Reddit ads unique. Accessed January 17, 2021. https://web.archive.org/web/202101

49. Investigators A of URP. The "All of Us" Research Program. *N Engl J Med*. 2019;381:668–76.

50. Homan C, Johar R, Liu T, *et al.* Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In*: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics; 2014:107-117.

51. Mowery D, Bryan C, Conway M. Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In*: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics; 2015:89-98.

52. Lai T, Shi Y, Du Z, *et al*. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics*. 2023;4:8–33.

53. Nashwan AJ, Abujaber AA, Choudry H. Embracing the future of physician-patient communication: GPT-4 in gastroenterology. *Gastroenterol Endosc*. 2023;1:132–5.

54. Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al.* Large language models in medicine. *Nat Med*. 2023;29:1930–40.

55. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*. 2023;90:104512.

Chapter 6

Artificial Intelligence–Enabled Analysis of Statin-Related Topics and Sentiments on Social Media

Sulaiman Somani, Marieke M. van Buchem, Ashish Sarraju, Tina Hernandez-Boussard & Fatima Rodriguez

JAMA Network Open, 24 April 2023

6.1 Abstract

Importance Despite compelling evidence that statins are safe, are generally well tolerated, and reduce cardiovascular events, statins are underused even in patients with the highest risk. Social media may provide contemporary insights into public perceptions about statins.

Objective To characterize and classify public perceptions about statins that were gleaned from more than a decade of statin-related discussions on Reddit, a widely used social media platform.

Design, Setting, and Participants This qualitative study analyzed all statin-related discussions on the social media platform that were dated between January 1, 2009, and July 12, 2022. Statin- and cholesterol-focused communities, were identified to create a list of statin-related discussions. An artificial intelligence (AI) pipeline was developed to cluster these discussions into specific topics and overarching thematic groups. The pipeline consisted of a semisupervised natural language processing model (BERT [Bidirectional Encoder Representations from Transformers]), a dimensionality reduction technique, and a clustering algorithm. The sentiment for each discussion was labeled as positive, neutral, or negative using a pretrained BERT model.

Exposures Statin-related posts and comments containing the terms statin and cholesterol.

Main Outcomes and Measures Statin-related topics and thematic groups.

Results A total of 10 233 unique statin-related discussions (961 posts and 9272 comments) from 5188 unique authors were identified. The number of statin-related discussions increased by a mean (SD) of 32.9% (41.1%) per year. A total of 100 discussion topics were identified and were classified into 6 overarching thematic groups: (1) ketogenic diets, diabetes, supplements, and statins; (2) statin adverse effects; (3) statin hesitancy; (4) clinical trial appraisals; (5) pharmaceutical industry bias and statins; and (6) red yeast rice and statins.

The sentiment analysis revealed that most discussions had a neutral (66.6%) or negative (30.8%) sentiment.

Conclusions and Relevance Results of this study demonstrated the potential of an AI approach to analyze large, contemporary, publicly available social media data and generate insights into public perceptions about statins. This information may help guide strategies for addressing barriers to statin use and adherence.

6.2 Introduction

Statins conclusively reduce morbidity and mortality from atherosclerotic cardiovascular disease (ASCVD), the number one cause of death worldwide[1,2]. Thus, statins are a cornerstone of the treatment and prevention of ASCVD across contemporary clinical practice guidelines[3]. Statins are one of the most commonly prescribed medications in the US, making up nearly 17% of all prescription pharmaceuticals in the past decade, in part because of their widespread availability, low cost, high degree of effectiveness, and breadth of clinical indications[4,5]. However, despite their well-established benefits and safety, statin use remains suboptimal in high-risk individuals with guideline-recommended clinical indications[6].

Understanding the reasons for statin underuse, including patient-level perspectives, is crucial to guiding public health and implementation efforts toward ASCVD prevention and treatment. While barriers to statin use have been explored using targeted surveys and focus groups, these may not be widely generalizable[7-9]. Social media platforms have become a promising avenue to understand and glean public views on health outside of the health care setting[10,11]. Such platforms are used for sharing personal stories, soliciting informal opinions, or sparking discussions about any topics. These platforms can rapidly disseminate information or misinformation through noncurated, transparent peer discussions[12]. For example, a study analyzed more than 11 000 social media posts and manually annotated them to understand belief patterns regarding statins, many of which reflected established concerns about their use, even in statin nonusers[13].

A social media platform that has steadily gained popularity for health-related purposes over the past years is Reddit, a discussion-based platform whose users can post questions, comments, and topics on a wide range of areas, including statins. The platform is free, has 52 million daily active users and approximately 430 million monthly users, and gets more than 30 billion views every month[14]. Given its widespread use, this platform may provide large-scale data on patient

views on statins that could be analyzed for novel insights and for misinformation that may affect statin adherence[15].

Manual analysis of social media data has limited feasibility given the speed and volume with which these data evolve over time. Artificial intelligence (AI) methods may facilitate the analysis and interpretation of these valuable patient-generated data that may influence health behaviors. Natural language processing (NLP) is a form of AI that ingests and interprets large volumes of textual data to glean insights and that can make actionable predictions on new data[16,17]. This study aimed to characterize and classify public perceptions about statins gleaned from more than a decade of statin-related discussions on Reddit.

6.3 Methods

The Stanford University Institutional Review Board deemed this qualitative study exempt from ethical review and the requirement for informed consent since it did not involve human participants. We followed the Standards for Reporting Qualitative Research (SRQR) reporting guideline[18].

Data Set and Search

Reddit was used as the data source for this study[19]. Data were collected between January 1, 2009, and July 12, 2022. The social media platform is composed of different communities, which use the r/ prefix and are focused on specific topics (e.g., r/gaming, r/worldnews, r/keto, and r/statins). Users may interact with the platform by creating a post to initiate a new discussion thread or by commenting on other users' posts as part of discussions. Most communities, including all posts and comments, are openly or publicly accessible and visible and thus require no account registration with the social media platform.

To create a list of statin-related discussions on this social media platform, we identified relevant communities by entering the words statin and cholesterol in the platform search engine and keeping those communities that were recommended by the search engine for both words. Across these communities, we

used an application programming interface called Pushshift to search all of the posts and comments for case-insensitive matching on the word statin and the generic or brand names for specific statins: atorvastatin, lipitor, rosuvastatin, crestor, pitavastatin, livalo, zypitamag, simvastatin, zocor, pravastatin, pravachol, lovastatin, altoprev, fluvastatin, and lescol (eFigure 1 in Supplement 1) [20].

Topic Modeling

Raw text that was collected from the social media platform was preprocessed to prepare it for automatic analysis. We used BERTopic, a state-of-the-art NLP technique that leverages the strength of BERT (Bidirectional Encoder Representations from Transformers) models to perform topic modeling to identify topics of discussion about statins. Briefly, BERTopic first embeds documents using a sentence-level BERT model, called Sentence-BERT, and applies an unsupervised machine learning technique called UMAP (Uniform Manifold Approximation and Projection) to simplify this representation [21]. The all-MiniLM-L6-v2 pretrained model was specifically chosen for the data set given its applicability for the social media platform and scientific content since it was already trained on more than 600 million posts and S2ORC, a data set containing more than 12.8 million papers in the field of medicine, among many other language data sets [22].

Topics were identified by spectral clustering, an algorithm for grouping similar discussions together into topics. Clustering performance was measured using 2 metrics: Silhouette coefficient and Davies-Bouldin index [23,24]. The Silhouette coefficient measures the similarity of a discussion to its own topic (cohesion) compared with its similarity to other topics (separation). A Silhouette coefficient (range, –1 to 1) that is closer to 1 indicates better performance. The Davies-Bouldin index takes a more global approach and compares the average similarity of each topic to its most similar topic. A Davies-Bouldin index score (range, 0-1) that is closer to 0 indicates more dispersed and less similar topics.

Since these topics can be granular and may overlap substantially, we performed a subsequent clustering analysis on a mathematical representation of these topics to find overarching themes of discussion (groups). Sensitivity analyses that maximized the Silhouette coefficient and Davies-Bouldin index were performed

to identify the optimal number of groups. Full details on preprocessing, model selection, dimensionality reduction (UMAP), topic and group clustering, and sensitivity analyses are provided in the eMethods in Supplement 1.

Sentiment Analysis

Sentiment analysis is a technique for identifying and extracting subjective information from text documents. A common form of sentiment analysis is to classify the tone of text documents into distinct categories, such as positive (e.g., "I love statins!") or negative (e.g., "I hate statins!") [25-27].

To assess the sentiments for each post, we used a pretrained BERT model, called RoBERTa, that was trained on social media posts [21]. RoBERTa offers multiclass labels (i.e., positive, neutral, or negative classification of text) and has been used in recent studies investigating health care problems using data from social media [28-31]. To quantify how sentiments varied across topics and groups, we transformed sentiment labels to scores: from negative to –1, neutral to 0, and positive to 1 [32]. Full details on algorithm and model choice, input data handling, and output transformation are provided in the eMethods in Supplement 1.

Statistical Analysis

We described discussion characteristics using mean and SD. Data analysis was performed from July to August 2022. Analysis was performed using the Python programming language, version 3.7.3 (Python Software Foundation) and multiple key libraries: scikit-learn, version 1.1.1; BERTopic, version 0.11.0; transformers, version 4.20.1; and matplotlib, version 3.5.2. Code that was developed for topic modeling and analysis is available on https://www.github.com/sssomani/statins_reddit.

6.4 Results

A total of 19 communities that contained both search terms statin and cholesterol were identified from a candidate list of 75 communities. From these 19 communities, a total of 10 233 unique statin-related posts and comments were

curated, which included 961 unique statin-related posts and 9272 unique statin-related comments from 5188 unique authors (Table 1). Posts were longer in mean (SD) number of characters than comments (1792.9 [2693.5] vs 839.9 [1122.1]). Most comments and posts were retrieved from matching the word statin (74.0%). Lipitor (12.7%) and atorvastatin (4.2%) were the second and third most common search terms, followed by Crestor (3.8%), simvastatin (1.8%), and rosuvastatin (1.7%). The communities that most frequently contained these posts and comments were r/keto (23.1%) and r/Cholesterol (21.3%), followed by r/diabetes (8.6%) and r/science (6.9%).

A total of 779 unique users (81.1%) authored all posts, and 4700 unique users (50.7%) authored all comments (Table 1). Most authors had between 1 and 5 posts (94.6%) (eFigure 2 in Supplement 1). The number of statin-related discussions increased by a mean (SD) of 32.9% (41.1%) per year (Figure 1A). However, the number of statin-related discussions on certain communities increased yearly (eg, r/Supplements, r/conspiracy, and r/diabetes) but decreased in others (such as r/Paleo and r/skeptic) (Figure 1B).

Table 1: Post and Comment Summary Statistics.

Characteristic	Category	All Discussions N (%)	Comments N (%)	Posts N (%)
Number of Discussions Scraped, N		10233	9272	961
Number of Characters, Mean (SD)		929.4 (1377.9)	839.9 (1122.1)	1792.9 (2693.5)
Unique authors		5188 (50.7)	4700 (50.7)	779 (81.1)
Search Word	statin	7571 (74.0)	6891 (74.3)	680 (70.8)
	lipitor	1295 (12.7)	1218 (13.1)	77 (8.0)
	atorvastatin	434 (4.2)	367 (4.0)	67 (7.0)
	crestor	389 (3.8)	357 (3.9)	32 (3.3)
	simvastatin	181 (1.8)	138 (1.5)	43 (4.5)
	rosuvastatin	177 (1.7)	141 (1.5)	36 (3.7)
	lovastatin	73 (0.7)	65 (0.7)	8 (0.8)
	pravastatin	71 (0.7)	56 (0.6)	15 (1.6)
	zocor	21 (0.2)	19 (0.2)	2 (0.2)

Table 1: (*Continued*)

Characteristic	Category	All Discussions N (%)	Comments N (%)	Posts N (%)
	pitavastatin	10 (0.1)	10 (0.1)	0 (0)
	fluvastatin	5 (0.0)	5 (0.1)	0 (0)
	livalo	4 (0.0)	3 (0.0)	1 (0.1)
	pravachol	2 (0.0)	2 (0.0)	0 (0)
Subreddit	keto	2364 (23.1)	2051 (22.1)	313 (32.6)
	Cholesterol	2182 (21.3)	1875 (20.2)	307 (31.9)
	diabetes	879 (8.6)	748 (8.1)	131 (13.6)
	science	706 (6.9)	704 (7.6)	2 (0.2)
	ketoscience	560 (5.5)	504 (5.4)	56 (5.8)
	nutrition	503 (4.9)	486 (5.2)	17 (1.8)
	ScientificNutrition	446 (4.4)	426 (4.6)	20 (2.1)
	news	443 (4.3)	443 (4.8)	0 (0)
	todayilearned	381 (3.7)	380 (4.1)	1 (0.1)
	conspiracy	375 (3.7)	355 (3.8)	20 (2.1)
	Supplements	367 (3.6)	326 (3.5)	41 (4.3)
	Health	238 (2.3)	230 (2.5)	8 (0.8)
	PlantBasedDiet	230 (2.2)	208 (2.2)	22 (2.3)
	askscience	165 (1.6)	154 (1.7)	11 (1.1)
	COVID19	136 (1.3)	135 (1.5)	1 (0.1)
	Paleo	112 (1.1)	105 (1.1)	7 (0.7)
	longevity	75 (0.7)	73 (0.8)	2 (0.2)
	skeptic	68 (0.7)	68 (0.7)	0 (0)
	stopusingstatins	3 (0.0)	1 (0.0)	2 (0.2)

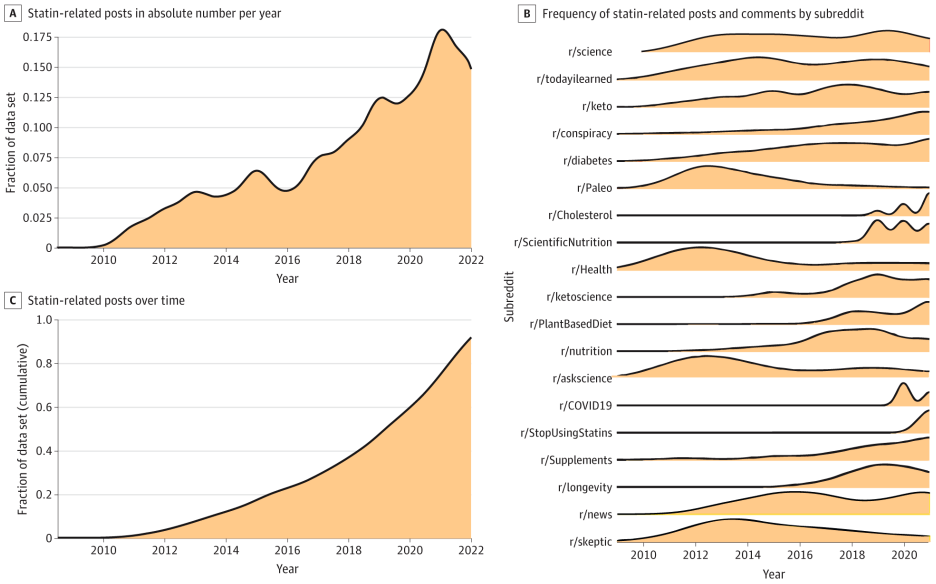


Figure 1: Statin-Related Posts and Comments Over Time. When mapped over time, the number of statin-related posts in the data set increased in absolute number per year (A) and cumulatively over time (C). The frequency of posts and comments varied by community over time (B).

A total of 100 topics of statin-related discussions were identified from the data set (eTable 1 in Supplement 1), with a performance Silhouette coefficient of 0.013 and a Davies-Bouldin index of 4.27. The 3 most common topics were elevated low-density lipoprotein cholesterol (LDL-C) when on a ketogenic diet (topic 1), advice and statin experience solicitation with changes in lipid panels (topic 2), and anecdotal perspectives on statin efficacy and adverse effects (topic 3). Other topics included adverse effects from statins (eg, topics 33, 39, 43, and 44); statin trial data and possible industry bias in their outcomes (eg, topics 22, 34, and 68); lifestyle alternatives, such as supplements (eg, topic 84); red yeast rice (eg, topic 69); dietary changes (eg, topics 5 and 65); improving outcomes with COVID-19 (eg, topics 63, 78, and 95); the interplay between coronary artery calcium scores and role of statins (eg, topic 67); and attitudes about statin development that used fetal stem cells (eg, topic 80). Hierarchical representation of these topics is shown in Figure 2A. Changes over time in the number of discussions per topic are presented in eFigure 3 in Supplement 1.

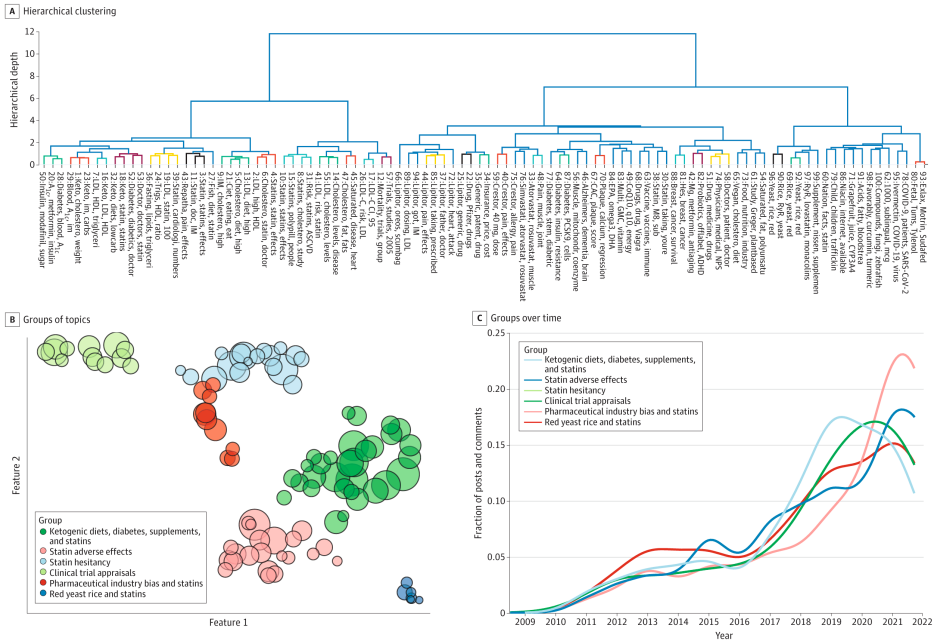


Figure 2: Topic modeling. Hierarchical (A) and spatial (B) representations of the 100 extracted topics (columns in A; circles in B) and 6 overarching groups are shown. In panel A, the y-axis represents the depth of the hierarchical tree corresponding to each node in the tree. The size of each topic represents the relative number of discussions grouped in that topic. Changes in the number of discussions for each of the 6 groups over time are shown in panel C. The x- and y-axes in panel B represent the 2 Uniform Manifold Approximation and Projection axes that were dimensionally reduced to allow for topic visualization. ADHD indicates attention deficit/hyperactivity disorder; ASCVD, atherosclerotic cardiovascular disease; CAC, coronary artery calcium; coQ10, coenzyme Q10; CYP3A4, cytochrome P450 3A4; DHA, docosahexaenoic acid; doc, doctor; EPA, eicosapentaenoic acid; HDL, high-density lipoprotein; IM, I am; hes, he is; LDL, low-density lipoprotein; LDL-C, low-density lipoprotein cholesterol; M8, mate; NPS, nurse practitioner; RyR, ryanodine receptor; trigs, triglycerides.

Six overarching groups (Figure 2B) from these 100 topics were identified after sensitivity analysis, maximizing the Silhouette coefficient and Davies-Bouldin index (eFigure 4 in Supplement 1). These groups were (1) ketogenic diets, diabetes, supplements, and statins; (2) statin adverse effects; (3) statin hesitancy; (4) clinical trial appraisals; (5) pharmaceutical industry bias and statins; and (6) red yeast rice and statins (Table 2). Temporal patterns in these thematic groups are shown in Figure 2C.

Table 2: Overview of Groups of Topics With Example Text.

Group	Topics	Posts (#)	Comments (#)	Description	Examples
1	1, 2, 5, 6, 7, 9, 13, 16, 18, 19, 20, 21, 23, 24, 27, 28, 29, 32, 36, 42, 50, 52, 53, 58, 67, 73, 77, 81, 83, 88, 91	691	3497	**Ketogenic Diets, Diabetes, Supplements, and Statins**	"Why is my cholesterol so high when I lost weight was working out had no fast food and my BP is good. I've seen it as low as 96/70 but it varies from 105/70 TO 120/80 while at home. Do you think I can fix this naturally?" "The rise of the calcium scan should make the decision easier. If you have a low calcium score then you probably don't need it. The higher the score the more obvious you do nwstatin because your cholesterol is calcifying. Also, baseline scans will help know if you are laying down an unusual amount of calcified plaque each year which would also argue for long term statin use." "High LDL - 9 months of Keto. My doctor is eager for me to go back on statin medication. I stopped taking it in 2019 due to side effects. I started keto in October 2019 and am now 45lbs lighter (34yo male). In October 2019 my number were: LDL: 181 HDL: 33 Triglycerides: 332 In June 2020 my numbers are: LDL: 256 HDL: 40 Triglycerides: 107 My doctor is extremely concerned about the LDL and says it must come down. There is a lot of conflicting information out there regarding keto and cholesterol. I'm hoping to receive some guidance and sources from the gurus in the subject."

Table 2: (*Continued*)

Group	Topics	Posts (#)	Comments (#)	Description	Examples
2	3, 11, 30, 33, 37, 39, 40, 43, 44, 48, 59, 60, 62, 66, 75, 76, 80, 85, 89, 93, 94, 96	102	1768	**Side Effects from Statins**	"I will never take a statin again." "That's very consistent with statins. Top 3 adverse events with the class are muscle aches (myalgia), joint aches (arthlagia), and fatigue. Call your doctor and see what they say. Typically, the recommendation is to take a 2-week holiday and then try a different statin at a lower dose." "My endocrinologist wants me to go on Lipitor but after reading info with the prescription, muscle loss is a side effect. I have enough trouble meeting my protein goals as it is and take supplemental Mg and Potassium to stave off muscle cramps, so its going to stay on the shelf for now."
3	4, 8, 10, 15, 31, 38, 41, 46, 49, 56, 63, 64, 70, 72, 86, 87, 95, 98	79	1775	**Statin Hesitancy: Initiation, Maintenance, and Alternatives**	"I'm not completely anti-statin, but I'm living proof that diet alone *can* change cholesterol drastically (see post above, or below if it got downvoted)." "...I'm just mad at the fact you said statins aren't healthy which is completely false in people using it for primary prevention or for diabetics..." "Given the sheer density of research performed on this remarkable spice [curcumin], it is no wonder that a growing number of studies have concluded that it compares favorably to a variety of conventional medications, including: * Lipitor/Atorvastatin (cholesterol medication)..."

Table 2: (*Continued*)

Group	Topics	Posts (#)	Comments (#)	Description	Examples
4	12, 14, 17, 26, 45, 47, 54, 55, 57, 61, 65, 78, 84	73	1271	**Clinical Trial Appraisals**	"Large study finds lower risk of death from COVID-19 in statin users. Probably because the Statins will kill you first." "People seem to think that statins are there to lower LDL-C. However, there are no trials to show statins work in that setting. statins don't really lower CVD events that much it's very controversial whether they lower all-cause mortality at all..." "Statins are lipid-lowering therapeutics with favorable anti-inflammatory profiles and have been proposed as an adjunct therapy for COVID-19. However, statins may increase the risk of SARS-CoV-2 viral entry by inducing ACE2 expression."

Table 2: (*Continued*)

Group	Topics	Posts (#)	Comments (#)	Description	Examples
5	22, 25, 34, 35, 51, 68, 71, 74, 79, 82	10	837	**Pharmaceutical Industry Bias and Statins**	"The cost of drugs truly is crazy. I work in a pharmacy and look at the cash cost of a lot of things just out of curiosity. Atorvastatin (generic lipitor) at a moderate dosage is like $450/month cash price at my pharmacy." "Uh oh, your cholesterol is high. Here's some Lipitor...and so on forever. Pharmaceutical companies are not charitable. They are corporations and they are in the business of symptom maintenance. Nothing more. No cures, only once a day pills. Doctors are not charities. They are businesses, and patients are customers." "You might not find this, but there's a lot of statins that can be taken with grapefruit juice."
6	69, 90, 92, 97, 99, 100	6	124	**Red Yeast Rice and Statins**	"red yeast rice is a statin, and statins should be avoided at all costs. If you have high cholesterol and also mind your diet, you may want to check LMHR communities and the cholesterol code. statins are basically mycotoxins and deplete you if fat soluble nutrients, like coQ10, vit D, K, A and E, and in all likelihood through these depletions worsen cardiovascular health." "When they discovered statin drugs for cholesterol, you look back at ancient Chinese remedies for cardiovascular health and find fermented red yeast rice over 1000 years ago which can actually create statin compounds."

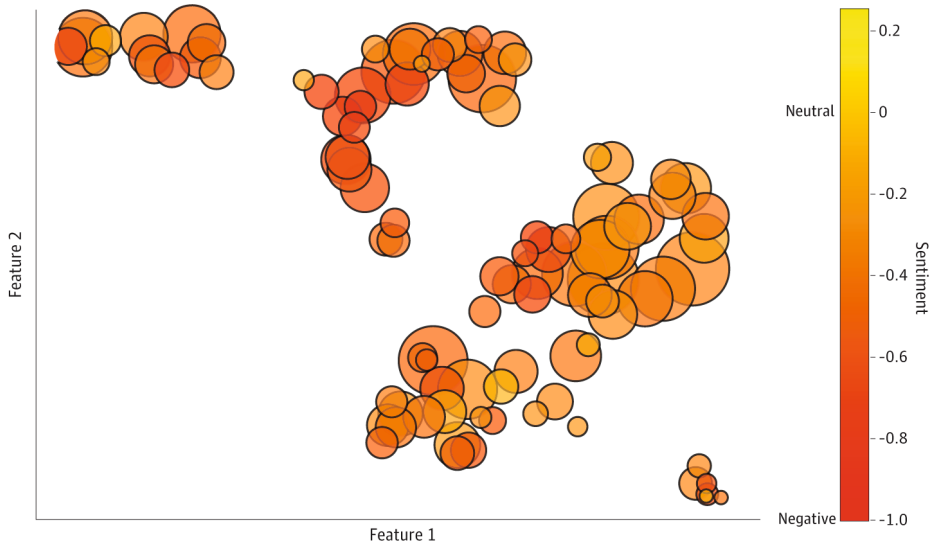


Figure 3: Sentiment analysis. Mean sentiment (color) across topics (circles) is shown. The size of each topic represents the relative number of discussions grouped in that topic. Mean sentiment scores that were close to –1 reflected a predominantly negative sentiment (red), close to 0 reflected an overall neutral sentiment (yellow), and close to 1 reflected an overall positive sentiment. Because no topics had a positive sentiment, the color map was truncated at 0.2 to allow for differentiation between negative and neutral sentiments. The x- and y-axes represent the 2 Uniform Manifold Approximation and Projection axes that were dimensionally reduced to allow for topic visualization.

Of the 10 233 discussions, 3151 had a negative (30.8%), 6815 had a neutral (66.6%), and 267 had a positive (2.6%) sentiment. Examples of discussions with a highly negative sentiment included the following: "So take a statin to decrease CVD risk ... but increase Alzheimer's risk? What a mess!," "And then statin kills you instead of covid," and "Statins are poison. It's a myth made up by the pharma industry obviously to make more $$$...." Examples of discussions with a highly positive sentiment included, "I love taking my simvastatin ... it also cured my psoriasis," "I sure do love me a statin," and "'I love Crestor, the taste is better." The mean (SD) sentiment score across all discussions was considered to be neutral

to negative (score, –0.28 [0.50]). Among topics, 2 had sentiment scores that were considered to be neutral (topics 62 and 100); all other topics had negative sentiments (Figure 3). None of the 6 groups represented a positive sentiment. Mean (SD) sentiment score by communities ranged from –0.52 (0.51) for r/conspiracy, which reflected a more negative sentiment, to –0.12 (0.35) for r/COVID19, which reflected a more neutral sentiment (eTable 2 in Supplement 1).

6.5 Discussion

This qualitative study leveraged more than a decade of patient-generated data from a social media platform to uncover public beliefs and perceptions about statins. Artificial intelligence methods were used to analyze 10 233 unique discussions from 5188 unique authors, which increased over time. This approach identified 100 topics from the discussion that represented 6 thematic groups: (1) ketogenic diets, diabetes, supplements, and statins; (2) statin adverse effects; (3) statin hesitancy; (4) clinical trial appraisals; (5) pharmaceutical industry bias and statins; and (6) red yeast rice and statins. Sentiment analysis demonstrated that these discussions had a predominantly neutral to negative sentiment. These findings highlighted community perceptions and potentially modifiable barriers to statin use.

This study demonstrated the potential of AI to automate the extraction and analysis of social media data to understand public perceptions on statins. It complements and extends prior work that evaluated statin attitudes and beliefs using manual qualitative analyses of Twitter posts[13]. Since topic prespecification can miss unexpected emerging ideas, the study's AI-enabled algorithm semiautomatically organized discussions into topics and broader groups while categorizing the sentiments on these discussions. By efficiently extracting and interpreting large volumes of valuable social media data, AI offers the prospect of monitoring public sentiment continuously at scale.

The primary groups of discussion in this study align with findings in prior studies that assessed patient beliefs about statins. For example, the USAGE (Under-

standing Statin Use in America and Gaps in Patient Education) study and PALM (Patient and Provider Assessment of Lipid Management) Registry have identified patient perceptions and barriers to statin adherence[6,7,33,34]. The present study corroborated some of these previous findings, uncovering similar reasons for statin hesitancy, including adverse effect profiles (e.g., myalgias, increased risk of diabetes, and cognitive dysfunction; group 2), disbelief in the LDL-C hypothesis, preference for lifestyle alternatives (e.g., dietary or supplementary modifications; groups 1, 3, and 6), and general disenfranchisement with health care (group 5). Furthermore, novel points of discourse were uncovered, such as the role of statins in improving COVID-19 outcomes, controversy regarding lifestyle improvement on a ketogenic diet but developing asymptomatic dyslipidemia, the use of the coronary artery calcium score as a convincing data point for initiating a statin, and hesitancy toward statins because of concerns that statins were made with fetal stem cells. Leveraging anonymous, noncurated, informal peer-to-peer discussions from public social media platforms, such as Reddit, may reveal topics and sentiments that may not be identified in formal targeted surveys or focus groups, clinical encounters, or clinical trial settings.

Sentiment analysis of social media posts and discussions about statins revealed a predominantly neutral to negative sentiment. Prior work has highlighted that bad publicity surrounding statins can affect patients' medication adherence. For example, a Danish study found that unfavorable media coverage about statins was associated with a decrease in new statin users and an increase in statin discontinuation among current statin users within 1 year of the press coverage[35]. More active public health efforts are needed to monitor health-related misinformation on readily accessible social media platforms.

Several examples of statin-related misinformation were identified, including distrust of the hypothesis that LDL-C has a causal association with heart disease (e.g., "I think LDL is pretty much irrelevant. Your HDL and Triglycerides are far more important" [r/keto, topic 7]) and of the association between COVID-19 and statins (e.g., "results imply the potential benefits of statin therapy in hospitalized subjects with COVID-19" [r/COVID19, topic 78]). While support for natural supplemental alternatives (e.g., "Red yeast rice is a statin basically, by

the way" [r/Cholesterol, topic 69]; "statins are basically mycotoxins and deplete you if [sic] fat soluble nutrients, like coQ10, vit D, K, A and E, and in all likelihood through these depletions worsen cardiovascular health" [r/Supplements, topic 90]) was powerful in this study across multiple thematic groups (1, 2, and 4), the recent SPORT (Supplements, Placebo, or Rosuvastatin) trial demonstrated a substantial decrease in LDL-C with low-dose rosuvastatin but not with common supplements compared with placebo[36]. Such misperceptions posted on social media platforms can serve as seeds for the spread of misinformation. For example, discussions on the social media platform used in this study are easily accessible since the platform does not require users to have an account to see content and are highly presented on search engines, with some individuals even using this platform as their default search engine[37]. These factors may allow the spread of misinformation to susceptible cohorts, as reported in a study of misinformation spread during the COVID-19 pandemic[38]. Prioritizing the understanding of and designing solutions for such health misinformation in the age of an infodemic was recently highlighted as a major priority for the Office of the Surgeon General[39], underscoring the importance of the current study.

Limitations

This study should be interpreted in the context of its limitations. First, spelling errors can lead to the mislabeling of discussions as having a false-positive sentiment for being associated with statins (e.g., creator misspelled as crestor, or colloquialization of the phrase stating facts as statin facts) or a false-negative sentiment (e.g., atorvastatin misspelled as atovrastatin). Second, user anonymity on the social media platform used in this study limits our knowledge of the demographic characteristics of the post and comment authors in this study, although social media users have been generally characterized as young in age (between 18 and 29 years) [40], which may inform the content of discussions. Third, statin-related content in social media discussions is visible for free to any patient looking for statin information on the internet. The data set used in this study was created from prespecified communities that were most associated with statin-related posts; thus, it may not include other statin-related posts and

comments on the social media platform. While growth of statin-related content was more than 30% year on year, growth statistics on the relevant communities before 2019 were not available; thus, we could not contextualize this growth with that of the data set. Fourth, the clustering techniques used in this study may have lumped together categories that were not intuitively similar or that were not easily clinically interpretable. This limitation points to how AI approaches can assist researchers in augmenting, but not replacing, the interpretation of big streams of data.

Conclusion

In this qualitative study, an AI method was developed to classify statin-related content on a social media platform into topics of discussion. These 100 topics were then organized into 6 thematic groups: ketogenic diets, diabetes, supplements, and statins; statin adverse effects; statin hesitancy; clinical trial appraisals; pharmaceutical industry bias and statins; and red yeast rice and statins. Such an AI approach can be used to analyze large, contemporary social media data and generate insights into public perceptions about statins. This information may help guide strategies for addressing barriers to statin use and adherence.

Author contributions

Drs Somani and Hernandez-Boussard had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Dr Somani and Ms van Buchem contributed equally as co–first authors, and Drs Hernandez-Boussard and Rodriguez contributed equally as co–senior authors.

- Concept and design: All authors.
- Acquisition, analysis, or interpretation of data: All authors.
- Drafting of the manuscript: Somani, van Buchem.

- Critical revision of the manuscript for important intellectual content: All authors.
- Statistical analysis: Somani, van Buchem.
- Obtained funding: Hernandez-Boussard, Rodriguez.
- Administrative, technical, or material support: Somani, Hernandez-Boussard, Rodriguez.
- Supervision: Hernandez-Boussard, Rodriguez.

Conflicts of interest

Dr Rodriguez reported receiving personal fees from HealthPals, Novartis, Novo Nordisk, and AstraZeneca outside the submitted work. No other disclosures were reported.

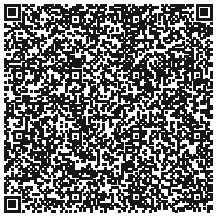
Funding

Dr Rodriguez was funded by grant 1K01HL144607 from the National Heart, Lung, and Blood Institute of the National Institutes of Health, a grant from the American Heart Association/Harold Amos Faculty Development Program, and grant 2022051 from the Doris Duke Foundation. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

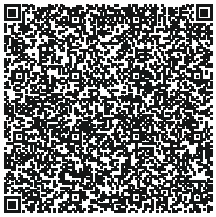
Data availability

See Supplement 2.

Supplement 1



Supplement 2



References

1. Baigent C, Keech A, Kearney PM, et al; Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet. 2005;366(9493):1267-1278.
2. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. Circulation. 2022;145(8):e153-e639.
3. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2019;73(24):e285-e350.
4. Preiss D, Tobert JA, Hovingh GK, Reith C. Lipid-modifying agents, from statins to PCSK9 inhibitors: JACC focus seminar. J Am Coll Cardiol. 2020;75(16):1945-1955.
5. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in prescription drug use among adults in the United States from 1999-2012. JAMA. 2015;314(17):1818-1831.
6. Bradley CK, Wang TY, Li S, et al. Patient-reported reasons for declining or discontinuing statin therapy: insights from the PALM Registry. J Am Heart Assoc. 2019;8(7):e011765.
7. Nanna MG, Navar AM, Zakroysky P, et al. Association of patient perceptions of cardiovascular risk and beliefs on statin drugs with racial differences in statin use: insights from the Patient and Provider Assessment of Lipid Management Registry. JAMA Cardiol. 2018;3(8):739-748.
8. Pokharel Y, Gosch K, Nambi V, et al. Practice-level variation in statin use among patients with diabetes: insights from the PINNACLE Registry. J Am Coll Cardiol. 2016;68(12):1368-1369.
9. Statin-intolerance registry. ClinicalTrials.gov identifier: NCT04975594. Accessed August 24, 2022. https://clinicaltrials.gov/ct2/show/NCT04975594
10. Forgie EME, Lai H, Cao B, Stroulia E, Greenshaw AJ, Goez H. Social media and the transformation of the physician-patient relationship: viewpoint. J Med Internet Res. 2021;23(12):e25230.
11. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res. 2016;16(1):442.
12. van der Linden S. Misinformation: susceptibility, spread, and interventions to immunize the public. Nat Med. 2022;28(3):460-467.
13. Golder S, O'Connor K, Hennessy S, Gross R, Gonzalez-Hernandez G. Assessment of beliefs and attitudes about statins posted on Twitter: a qualitative study. JAMA Netw Open. 2020;3(6):e208953.
14. Curry D. Reddit revenue and usage statistics (2022). October 2, 2020. Accessed August 14, 2022. https://www.businessofapps.com/data/reddit-statistics/
15. Gatewood J, Monks SL, Singletary CR, Vidrascu E, Moore JB. Social media in public health: strategies to distill, package, and disseminate public health research. J Public Health Manag Pract. 2020;26(5):489-492.
16. Alam KN, Khan MS, Dhruba AR, et al. Deep learning-based sentiment analysis of COVID-19 vaccination responses from Twitter data. Comput Math Methods Med. 2021;2021:4321131.

17. Sarker A, Gonzalez-Hernandez G, Ruan Y, Perrone J. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. JAMA Netw Open. 2019;2(11):e1914672.
18. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med. 2014;89(9):1245-1251.
19. Reddit website. Accessed August 24, 2022. https://www.reddit.com
20. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift Reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. 2020:14(1):830-839.
21. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-Networks. arXiv. Preprint posted online August 27, 2019.
22. Hugging Face. sentence-transformers/all-MiniLM-L6-v2. Accessed December 2, 2022. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
23. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65.
24. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;1(2):224-227.
25. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 2020:38-45.
26. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Sci China Technol Sci. 2020;63(10):1872-1897.
27. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7(2):e12239.
28. Kolluri N, Liu Y, Murthy D. COVID-19 misinformation detection: machine-learned solutions to the infodemic. JMIR Infodemiology. 2022;2(2):e38756.
29. Noraset T, Chatrinan K, Tawichsri T, Thaipisutikul T, Tuarob S. Language-agnostic deep learning framework for automatic monitoring of population-level mental health from social networks. J Biomed Inform. 2022;133:104145.
30. Baker W, Colditz JB, Dobbs PD, et al. Classification of Twitter vaping discourse using BERTweet: comparative deep learning study. JMIR Med Inform. 2022;10(7):e33678.
31. Anetta K, Horak A, Wojakowski W, Wita K, Jadczyk T. Deep learning analysis of Polish electronic health records for diagnosis prediction in patients with cardiovascular diseases. J Pers Med. 2022;12(6):869.
32. Hugging Face. J-hartmann/sentiment-roberta-large-english-3-classes. Accessed August 25, 2022. https://huggingface.co/j-hartmann/sentiment-roberta-large-english-3-classes
33. Cohen JD, Brinton EA, Ito MK, Jacobson TA. Understanding Statin Use in America and Gaps in Patient Education (USAGE): an internet-based survey of 10,138 current and former statin users. J Clin Lipidol. 2012;6(3):208-215.

34. Golder S, Weissenbacher D, O'Connor K, Hennessy S, Gross R, Hernandez GG. Patient-reported reasons for switching or discontinuing statin therapy: a mixed methods study using social media. Drug Saf. 2022;45(9):971-981.

Kriegbaum M, Liisberg KB, Wallach-Kildemoes H. Pattern of statin use changes following media coverage of its side effects. Patient Prefer Adherence. 2017;11:1151-1157.

Laffin LJ, Bruemmer D, Garcia M, et al. Comparative effects of low-dose rosuvastatin, placebo, and dietary supplements on lipids and inflammatory biomarkers. J Am Coll Cardiol. 2023;81(1):1-12.

Sullivan M. Is Reddit a better search engine than Google? February 17, 2022. Accessed December 3, 2022. https://www.fastcompany.com/90722739/is-reddit-a-better-search-engine-than-google

Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep. 2020;10(1):16598.

Office of the Surgeon General (OSG). Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment. US Department of Health and Human Services; 2021.

Stocking G, Holcomb J, Mitchell A. 1. Reddit news users more likely to be male, young and digital in their news preferences. February 25, 2016. Accessed October 26, 2022. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Part 2

Evaluating the Added Value in Clinical Practice

Chapter 7

The Added Value of the Artificial Intelligence Patient-Reported Experience Measure (AI-PREM Tool) in Clinical Practise: Deployment in a Vestibular Schwannoma Care Pathway

Olaf M. Neve, Marieke M. van Buchem, Marleen Kunneman, Peter Paul G. van Benthem, Hileen Boosman & Erik F. Hensen

PEC Innovation, 15 December 2023

7.1 Abstract

Objectives

Patient-reported experience measures (PREMs) can be used for the improvement of quality of care. In this study, the outcome of an open-ended question PREM combined with computer-assisted analysis is compared to the outcome of a closed-ended PREM questionnaire.

Methods

This survey study assessed the outcome of the open-ended questionnaire PREM and a close-ended question PREM of patients with unilateral vestibular schwannoma in a tertiary vestibular schwannoma expert centre.

Results

The open-ended questions PREM, consisting of five questions, was completed by 507 participants and resulted in 1508 positive and 171 negative comments, categorised into 27 clusters. The close-ended questions PREM results were mainly positive (overall experience graded as 8/10), but did not identify specific action points. Patients who gave high overall scores (>8) on the close-ended question provided points for improvement in the open-ended question PREM, which would have been missed using the close-ended questions only.

Conclusions

Compared to the close-ended question PREM, the open-ended question PREM provides more detailed and specific information about the patient experience in the vestibular schwannoma care pathway.

Innovation

Automated analysis of feedback with the open-ended question PREM revealed relevant insights and identified topics for targeted quality improvement, whereas the close-ended PREM did not.

7.2 Introduction

Patient experiences are important indicators of the quality of care. According to the national health service (NHS) policies, patient experiences reflect the compassion, dignity and respect for patients during health care delivery[1,2]. Moreover, these experiences may hold important insights for quality improvement[3]. Adequate tools to survey and analyse patient experiences are therefore essential. Patient experiences can be measured using patient-reported experience measures (PREMs), usually in the form of questionnaires[4].

PREMs may be considered subjective, but a positive association between PREM results and other quality domains has been reported[5]. PREM scores are positively but weakly associated with patient safety and clinical effectiveness, which suggests that improving patient experiences may enhance the overall quality of care[6,7]. Today, there are many different PREMs in use; most of them are disease or treatment specific and consist predominantly of closed-ended questions[8,9,10,11,12]. Some generic PREMs have been developed and are used to benchmark hospitals at a regional, national or international level[13,14,15,16,17].

The increased use of PREMs is incentivised by regulatory bodies in the United Kingdom and United States of America. Frequently, PREMs are collected and analysed but translating the results into changes in clinical practice remains challenging due to organizational, professional and data-related barriers[18,19,20]. The lack of a quality improvement infrastructure is one of these barriers[20]. Furthermore, patient experiences are not always adopted by clinicians, because the PREM results do not provide insights relevant to their daily workflow, or because the feedback is not specific enough to allow translation into concrete action points[3,19]. When PREM results are not translated into clear and actionable points of improvement for care providers, PREMs risk to be viewed as measurement for the sake of measurement rather than as valuable instruments for improving the underlying care[21].

In contrast to closed-ended questions that steer a patient's feedback to a specific topic, open-ended questions enable patients to provide feedback on all aspects of care that matter to them[22]. This feature makes open-ended questions more patient-centred and yields more specific information, facilitating concrete quality improvement measures[23]. However, the analysis of free-text answers is time-consuming and too laborious to use in clinical practice[23,24].

Artificial intelligence (AI) techniques are able to automatically detect the topics and sentiment of patients' free text comments and help identify actionable insights out of PREMs[25,26].

Currently used PREMs are not ideally suited for the full exploitation of the potential of AI-techniques. First, current questionnaires often contain questions with a sentiment comprised in the question itself. (e.g., 'what went remarkably well during your stay?' or 'what could we improve?'). In addition, questions such as these invite short, monosyllabic answers, which are difficult to categorize[25]. To tackle these problems several modifications to commonly used PREMs are needed. A new AI-PREM tool has been developed and validated by Van Buchem et al.[27], with open-ended generic questions (i.e., not targeted at a specific disease, care pathway, department or healthcare centre) and suited for computer analysis by removing the sentiment from the question. The questions were focused on the Picker dimensions of patient-centred care to reduce the number of topics in an answer (e.g., What did you think about the information provision?)[27].

The primary aim of this study was to determine the added value of the AI-PREM tool compared to a conventional PREM with respect to identification of actionable points for quality improvement. The secondary aim was to assess the influence of socio-demographic determinants on AI-PREM completion and results. To do so, we have deployed the AI-PREM in a vestibular schwannoma integrated practice unit (IPU) in a vestibular schwannoma expert centre in the Netherlands.

7.3 Methods

Context

Vestibular schwannomas are rare benign intracranial tumours, which typically cause hearing loss, tinnitus and balance disorders. A majority (52–78%) of the tumours is non-progressive. In these cases active surveillance with prolonged follow-up is usually the management strategy of choice[28]. In case of very large or progressive tumours, surgery or radiotherapy is indicated to prevent future complications such as brain stem compression. After active therapy, prolonged follow-up is warranted to detect residual or recurrent disease. Because of the long follow-up required (with or without active treatment) and near to normal life expectancy with adequate management of the tumour, patients with a vestibular schwannoma often accumulate extensive experience with healthcare professionals and centres.

Design

This descriptive case study evaluated the outcomes of an open-ended question PREM and a close-ended question PREM employed in a vestibular schwannoma IPU. A non-responder analysis was performed, the outcomes of both PREM were analysed, and the ceiling effect was evaluated in a direct comparison. In addition, the interpretation and the selection of actionable points of improvement by the IPU team based on these outcomes was observed. The process to come from PREM results to actionable points of improvement is reported.

The study was performed at the Leiden University Medical Centre, a tertiary university hospital, and expert centre for vestibular schwannomas in The Netherlands. At our centre, patient care is organized in an IPU, including otorhinolaryngologists, neurosurgeons, radiation oncologists and radiologists. The combination of chronic care and the multidisciplinary organization in an IPU are ideal to investigate the added value of AI-PREM for quality improvement.

Participants

This study was part of larger study on long term quality of life in vestibular schwannoma patients. For longitudinal follow-up patients who participated in

2014 in a cross-sectional survey on quality of life in vestibular schwannoma patients were re-invited for participation[29]. Using this patient group allowed the analysis of non-responders based on the data collected in 2014. In 2014, all consecutive patients who were diagnosed or treated for a unilateral vestibular schwannoma since 2003 at the IPU were eligible for inclusion. Patients under 18 years, patients with insufficient proficiency in the Dutch language to complete the questionnaires or patients with other skull base pathologies were excluded. Data collection took place between June and September 2020. The local medical research and ethics committee has waived the necessity for medical ethical approval under Dutch law and approved the study regarding data handling and privacy regulations (N19.112).

Data collection

After providing informed consent, patients were asked to complete two validated PREM questionnaires either electronically or on paper. First, participants completed the AI-PREM, consisting of five open-ended questions about their experiences with the care delivery[27]. The five questions (Box 1) addressed the following themes: information provision, personal approach, collaboration, organization and other experiences, and were based on the Picker dimensions of patient-centred care[13,30]. The free-text answers were analysed using natural language processing techniques, which divided the free-text answers into clusters of positive and negative comments. These techniques are described in more detail by Van Buchem et al.[27] The output of the AI-PREM are clusters of positive and negative comments for each of the five questions. The output was accessible in a easily intelligible dashboard. This dashboard was able to show the thematically clustered patient feedback, differentiate negative from positive clusters, and quantify the number of comments per thematic cluster. In addition, the IPU team could access the full individual patient comments the clusters were based on (as raw text).

Box 1

Questions AI-PREM [30]
Q1: How was the provided information?
Q2: How was the personal approach?
Q3: How was the collaboration between healthcare professionals?
Q4: How was the organization of care?
Q5: What else would you like to share about your experience?

Second, participants completed the Patient Experience Monitor (PEM) consisting of fifteen closed-ended questions about the patient's experience[14]; The PEM outcomes are proportions of patients which answered with a certain multiple choice option. For example, the proportion of the total number of respondents that trusted their physician fully.

Third, patients were asked to complete a disease-specific quality of life questionnaire of 26 items, the Penn Acoustic Neuroma Quality Of Life (PANQOL) [31,32]. Furthermore, demographic information (sex, age and education level) was acquired. Statistics Netherlands' (CBS) definition for low, middle and high education level was used, which follows the international standard classification of education[33].

Treatment modality, tumour size at baseline, and time since diagnosis were acquired from the electronic patient records. Treatment was coded as either active surveillance, surgery or radiotherapy. Tumour size was classified according to Kanzaki et al. as intrameatal, small, moderately large, large or giant tumour[34].

Statistical analysis

Statistical analyses were performed in R version 4.0.5 using Rstudio 1.3.959 (Rstudio, PBC, Boston).

For the demographics and non-responder analysis, means and standard deviation (sd) were calculated for normally distributed numerical variables, and medians and interquartile ranges (IQR) when not normally distributed. For cat-

egorical variables, percentages and frequencies were calculated. Demographics of non-responders, responders and one-word responders were compared using the unpaired t-test for continuous and chi-squared test for categorical variables. One word responders were defined as patients who provided a one-word answer for all open-ended questions (e.g., “well”, “fine”, or “bad”). Bonferroni correction for multiple testing was used to prevent type-I errors. Incomplete questionnaires were omitted in the analysis.

The ceiling effect, a well-known feature of PREMS, was analysed using the overall experience question of the PEM. In a separate analysis, the outcome of the AI-PREM was evaluated for patients who scored >8 out of 10 on the PEM questionnaire (i.e. provided overall very positive feedback). This analysis was used to assess the capability of the AI-PREM to identify feedback that could be used for quality improvement from patients that were overall positive about their experience with the IPU.

Intervention

The results of the AI-PREM and PEM were used to identify actionable point for quality improvement. The process to analyse, interpret and translate the results are described stepwise. First, results were analysed and placed in the local context by the IPU team. This team, consisting of a deputy of each medical specialism, a researcher, a case manager and supportive staff, used their knowledge of the IPU combined with the PREM results to select feasible and effective projects.

7.4 Results

In total, 536 patients provided informed consent resulting in a 62% response rate, as is shown in Fig. 1. Non-responders more often had a lower level of education (32% vs 44%) but a comparable mean age and male/female ratio to the responders, as shown in Table 1.

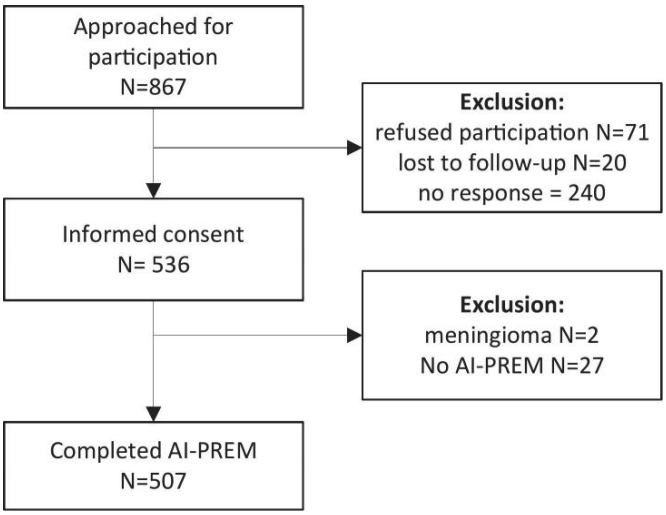


Figure 1: Flowchart study participants.

Compared to the population of vestibular schwannoma patients, the study population had a somewhat higher mean age (67.4 vs. 61.1 years) as a result of the long term follow-up. Also, the ratio of patients that received active intervention (radiotherapy or surgery) was higher (42% vs 51%), also as a result of the fact that they have been under observation for longer.

AI-PREM outcomes

The AI-PREM was completed by 507 patients, of whom 79 (16%) were one-word responders. As shown in Table 1, one-word responders were more often male, two years older and had a lower education level, but these differences were not statistically significant after correcting for multiple testing. A group of 27 patients did provide informed consent but did not complete the AI-PREM and two patients were excluded because of a pathology different to vestibular schwannoma.

Table 1: Baseline demographics.

	Non-responders	Not completed	Completed	One-word answers
	N= 331	N=28	N=507	N=79
Sex (male)	49%	50%	53%	65%
Age (sd)	68.0(12.3)	69.9 (10.5)	67.4 (11.0)	69.7 (9.6)
Education level				
Low	44%	44%	32%	41%
Middle	25%	33%	30%	27%
High	31%	22%	38%	33%
Treatment				
Observation	61%*	50%	46%	49%
Surgery	26%*	29%	38%	34%
Radiotherapy	13%*	14%	13%	16%
Quality of Life (sd)	69.8 (19.8)*	66.8 (15.5)	69.2 (18.1)	70.4 (17.3)

Demographics are shown for non-responders and responders. Both incomplete and completed questionnaires are shown. One-word responders are a subcategory of completed questionnaires, in which patients completed only one-word answers, such a "good" or "bad", on each open-ended question. Quality of life shows a disease-specific quality of life questionnaire ranging from 0-100. Higher scores indicate better quality of life. sd= standard deviation. *= data acquired in 2014

Table 2 shows the different feedback clusters of the five PREM questions including the number of comments per cluster and an example of a raw data comment. The majority of comments was classified as positive. All positive clusters contained many short or monosyllabic responses containing "well" or "fine", which did not provide additional information or context other than the subject of the question. Negative answers were in general more detailed and contained more words. Due to the diverse nature of the negative feedback, there were more thematic clusters, each containing less individual comments. For example, three negative clusters stated that personal approach was lacking (N = 3), limited (N = 3), or insufficient (N = 6). Another interesting finding was that different patients may experience certain aspects of care in a contradicting way. Therefore, the number of patients with a positive or a negative experience with the specific aspect of care was quantified, in order to put the feedback into perspective and help decide whether and which action should be taken to improve the IPU. For example, the number of patients who provided positive feedback on scheduling appointments on the same day (N = 8) outnumbered those who provided negative feedback on this topic (N = 2).

Table 2: AI-PREM clusters open-ended question answers.

Question	Information provision		Personal approach		Collaboration		Organisation		Other experiences	
Clusters	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*	*positive*	*negative*
	Well n=178	Limited* n=18	Well n=175	Insufficient n=6	Well n=215	Other hospital n=5	Well n=205	Appoint-ment* n=34	Well* n=105	Aftercare n=14
	Clear* n=178	Lengthy n=8	Fine n=55	Reserved n=3	Fine n=98	Communication n=7	Fine n=59	Reachability n=5	Positive n=8	Waiting time* n=21
			Pleasant conver-sation* n=101	Personal ap-proach n=3		Bad n=4	Well orga-nized* n=62			
			Personal ap-proach n=24	Limited n=3		Suboptimal n=3				
				Support` n=4		Better n=3				
				Experiences* n=5		Scan n=9				
				Attention n=7		Insufficient* n=8				
Leftovers	n=3	n=0	n=5	n=0	n=10	n=1	n=17	n=0	n=8	n=0
Example raw data quotes (from clus-ter with *)	"The information about the disease and symptoms was clear, informative and understandable."	"Limited. Several of my symptoms, that were in my opinion related to the tumour, were not addressed at all."	"Excellent. Understanding and sympathetic about the symp-toms. "	"The doctor's 'empathic capaci-ty' sometimes did not align with the patient's experi-ences/feelings."		"Scheduling a follow-up scan was sometimes difficult."	"Appoint-ments were mostly scheduled on the same day, which was pleasant."	"Sometimes you had to wait a long time for the appoint-ment, no appointments scheduled on the same day. Difficult to reach by phone."	"The vestib-ular schwan-noma team is well-coordi-nated."	"Long waiting time for the scan results"

PEM outcomes

The PEM was completed by 490 patients. In general, the patients completed the PEM very positively and the overall experience was graded with an 8 (±1.2 sd) on a 1 to 10 point scale. For example, 95% of the patients trusted their physician, and 93% indicated they had enough time to discuss their problem with the physician. Furthermore, 93% of patients said they discussed what to do after the consultation, and 89% said they were informed about their treatment's pros and cons. The majority (87%) found the physician's explanation understandable. Only 1% indicated they could not ask questions to their consulting physician.

The question with the most negative responses concerned the waiting time in the outpatient clinic. 21% of patients indicated they had to wait >15 min. Of this group, 10% would have preferred to receive more information about the estimated waiting time.

Comparison between PREMs

Table 3 shows the AI-PREM results of patients who scored an overall experience >8 out of 10 points on the PEM questionnaire. These patients had also rather positive experiences on the AI-PREM and only a limited number of negative comments. Still, these comments provided useful and detailed information about the IPU. For example, one patient stated: "I would have liked to hear about the treatment of vertigo with exercises sooner". Other patients mentioned: "There was some misunderstanding about by whom and when I was called about an appointment.", "The collaboration between hospitals was poor.", and "I was discharged from the hospital too soon and without instructions."

Table 3: AI-PREM results of patients with an overall PEM scores of >8/10.

	Negative		Neutral		Positive	
	count	%	count	%	count	%
Information provision	3	2%	37	23%	122	75%
Personal approach	2	1%	35	22%	125	77%
Collaboration	6	4%	35	22%	121	75%
Organisation	6	4%	35	22%	121	75%
Other experiences	7	4%	90	56%	65	40%

Observation of the interpretations of results

The results of the close-ended PEM questionnaire were predominantly positive, which was considered motivating information for the IPU team. However, for quality improvement these positive reactions could not be translated to action points for improvement. Conversely, the AI-PREM results provided more detailed information about the positive and negative experiences, even from patients that provided overall positive feedback. This information could be used to identify action points.

The process to identify action points for improvement is shown in Fig. 2. First, the IPU team analysed the results of the AI-PREM and explored the negative clusters of patients' experiences for potential quality improvements. The automated sentiment analysis and clustering of comments was used to identify topics of interest. These topics of interest were subsequently further explored by the IPU team through targeted evaluation of clustered patient comments (raw text). These raw texts were valued in the context of the IPU organization. When potential action points emerged they were discussed in the meeting and weighed against possible positive feedback regarding the same topic.

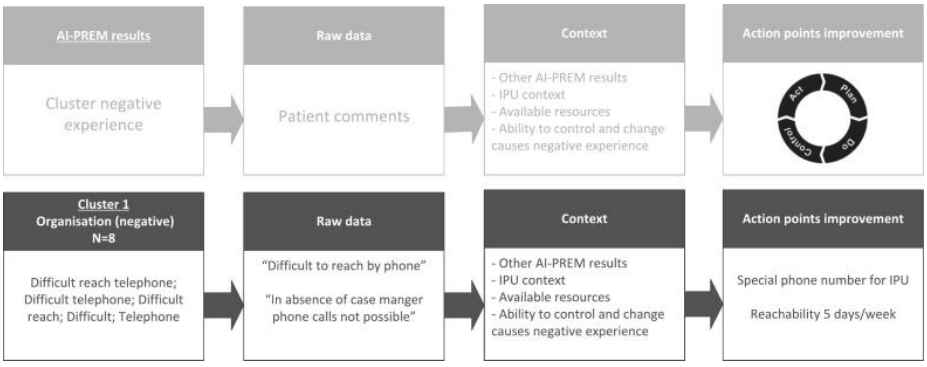


Figure 2: Process from AI-PREM results to quality improvement.

The process steps from using the AI-PREM results to identify action points for quality improvement are shown in grey. The second row shows the process steps of the identified action point reachability by phone.

In all, the IPU team selected three action points for quality improvement based on actionability, feasibility and number of patients sharing the particular (negative) experience. The chosen action points were improving the reachability by phone, reducing the time between the MRI and the consultation to discuss the result and improving the communication with referring hospitals.

7.5 Discussion and conclusion

To our knowledge, this is the first study in which a PREM with open-ended questions is directly compared to a traditional PREM with close-ended questions. Both questionnaires allowed evaluation of patient experiences with the care provided by the vestibular schwannoma care pathway. Both questionnaires reported overall positive patients' experiences.

The PEM enabled an easy and quick quantitative analysis of the overall experience. Most results showed ceiling effects and the predefined answer categories were less suited for identification of points of improvement, especially in the context of predominantly positive experiences. The AI-PREM seemed to have a greater potential to identify actionable points for quality improvement because of the broader focus and the more detailed descriptions, especially of negative experiences. With the AI-PREM, feedback with improvement points could be obtained even from patients with very positive experiences (as judged on the PEM scores).

An essential feature determining feasibility for clinical use was the automated analysis of the open text PREMs to reduce the workload. Still, the human component in the analysis is essential to interpret the algorithm's results and combine this with the clinical context of the IPU to translate the feedback into actionable points of improvement. Furthermore, the AI-PREM combined output of quantitative and more qualitative data. This combination of sentiment scores, the number of comments per cluster and a traceback to the individual reported experience facilitated decision making for quality improvement. In contrast,

the use of the structured PEM for identification of points of improvement was limited due to a small number of reported negative experiences.

The AI-PREM results showed that most comments were positive, but negative comments provided more detailed descriptions, including more context. Positive comments were more often one-word answers and generic. These findings were also described by Cunningham et al. while analysing almost 7000 open-text comments[22]. Positive comments are essential to put the negative ones into context and prioritize action points for improvement. For example, when many comments are positive about scheduling appointments, some negative comments on this cluster might be outliers, making this a less urgent target for quality improvement. In addition, positive comments can be used as motivators for the IPU team and can contribute to increasing patient safety following the Safety-II paradigm, which focuses on the things that go right rather than focusing on things that go wrong[16,35].

Other studies, focussing on patients narratives, have reported that the patients' comments on their experience with disease and care delivery generally provide mainly positive outcomes[16,17,36]. For example, the study of De Rosis et al. reported mainly positive comments which could be used for to identify positive aspects, which could be used for quality improvement by a 'learning by excellence' strategy. While this is valuable, learning by excellence in itself has a limited ability to to identify actionable points for improvement. The AI-PREM presented here has the ability to show and quantify positive comments but at the same time identify points of improvement, even in the feedback of patients that are overall positive about their experience in the IPU. In doing so, a more nuanced feedback of patients on the care delivery is made possible. While we find, like previous reports, that a large majority of patients provide positive comments, we were also able to extract actionable points of improvement even from patients with generally positive feedback.

Also in research settings, generic PREMs are used to evaluate the quality improvement targeted at improving the overall patient experience[36]. Improving organizational factors for a better patient experience will not only benefit pa-

tients but has also been shown to enhance physician satisfaction[37].However, achieving improvements in the patient experience can be challenging[38]. A large proportion of patients report high PREM scores. This ceiling effect might be caused by appreciation or social desirability bias[39,40]. In this study, the PEM results also show this ceiling effect, which is challenging from a quality improvement perspective since these already high scores can be hard to improve on. When trying to improve patient care, focussing on overall patient satisfaction or PREM scores may therefore be less effective than evaluating the negative comments in detail. Moreover, this study shows that even patients with a positive overall experience (as reported in the PEM) may still have feedback indicating points of improvement (identified with the AI-PREM). The AI-PREM design allows for an in-depth analysis of the comments by grouping them together in clusters based on sentiment and similar word content. Consequently, the actual remarks concerning a certain topic made by individual patients can be accessed, providing all necessary detail, without manually going through all questionnaires to extract information about the topic at hand. This approach, which yields both quantitative and qualitative data from free-text answers, saves time yet allows patients to comment freely on their experience with all aspects of care, detailed analysis of their feedback and identification of specific points of improvement.

A potential problem of using PREMs for quality improvements is a selection bias of the patients who complete the PREMs. When the responders are not a random sample of the total patient population the risk for inadequately aimed quality optimisations exists. Younger patients and black, indigenous and people of colour tend to report less positive patient experiences[41,42]. So it is important to include answers of these groups in the analysis for quality improvement. The non-responder analysis showed a larger proportion of lower education level in this group. There were no age differences, but one-word responders were on average slightly elder. These aspects should be considered when interpreting the PREM results to prevent nonresponse errors[43].

In addition, open-ended PREMs might reflect the a priori expectations and perceptions of care. When the provided care meets the expectations, patients might

not provide feedback but they probably will when the experience is worse or much better than their expectations. This phenomenon is especially important since different populations have different expectations of care delivery[44,45]. The evolution from patient satisfaction (e.g., how would you rate the information you received about your treatment?) towards experience (e.g., did you receive information about your treatment?) has mitigated the risk of such bias[45]. However, open-ended questions in structured PREMs are often focussed on patient satisfaction (e.g., "What went remarkably well during your stay?"). The AI-PREM questions focus more on the experience and reduce but not neutralize the risk of expectation bias.

In this study, a patient population was selected that had already participated in previous research. These dedicated participants might introduce some selection bias. When collecting the PREMs prospectively, the response rate might, therefore, be lower. Another limitation was the prolonged recall period since the last visit to the hospital in this research. The period exceeded the 4–6 weeks used in the PEM validation study[14]. This prolonged period might have limited the output of the PREMs[2]. However, the comparison between the two PREMS was not affected since both questionnaires were completed simultaneously.

Experiences of deployment in a vestibular schwannoma IPU

The IPU team used the PREM results to identify actionable points for quality improvement. This entailed a process of interpretation of the PREM results and analysing them in order to use them to improve clincal practice. Important parameters during the IPU team discussions were the quantitative results and the positive feedback clusters. The quantitative information (how many patients shared the same view) was useful in determining the extent of the problem. However, the positive feedback was essential too, for putting certain negative comments into perspective and prioritizing and focusing actions on improving the care delivery. Taking action based on the negative comments only could mistakenly alter aspects of care that provided a positive experience for most patients.

In addition, the potential of the IPU to improve or change the underlying causes of the negative experience was discussed. For example, a negative patient experience about a lack of parking space is beyond the control of the IPU, but the communication about the appointments is within the sphere of influence of the IPU. When potential action points were within the sphere of influence, the available resources needed to perform an improvement cycle were identified to see whether an improvement cycle was feasible. Finally, the IPU team decided to start a plan, do, check, act cycle.

Innovation

With the growing interest in patient-centeredness of care comes a growing need to adequately assess the patient experience with care delivery. The AI-PREM may be a tool that allows patients to freely comment on their experience yet is economic with the time and effort invested by healthcare professionals to analyse the feedback, although the time and effort invested by patients to complete the AI-PREM should also be considered. To make the efforts of patients worthwhile, PREMs should be used to improve care delivery, rather than as an administrative requirement. Future research should evaluate the applicability of the AI-PREM in different clinical settings. Because of the generic nature of the AI-PREM questionnaire, it seems likely to be of value in a multitude of different diseases, care pathways, or healthcare centres. In addition, the ability of the AI-PREM to detect longitudinal changes in the quality of care and/or the effect of measures to improve the quality of care may be the subject of future research.

Conclusion

Patient experiences are an essential aspect of quality of care. This study showed the added value of open-ended PREM questions in assessing patient experiences. The AI-PREM provided insights into both positive and negative experiences and allowed the detection of actionable targets for quality improvement in an IPU. Because of its automated analysis and readily accessible results, the evaluation of the patient experience with the vestibular schwannoma care pathway could be performed by IPU clinicians and translated into action points relevant to context of the clinical IPU.

Authorship contribution

O.M. Neve: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. M.M. van Buchem: Data curation, Investigation, Software, Writing – review & editing. M. Kunneman: Writing – review & editing. P.P.G. van Benthem: Writing – review & editing. H. Boosman: Conceptualization, Methodology, Writing – review & editing. E.F. Hensen: Conceptualization, Supervision, Writing – original draft.

References

1. Department of Health . In: High quality care for all: NHS next stage review final report. D.o. Health (Ed.), editor. The Stationery Office; London: 2008.
2. Manary M.P., Boulding W., Staelin R., Glickman S.W. The patient experience and health outcomes. N Engl J Med. 2013;368(3):201–203.
3. Gleeson H., Calderon A., Swami V., Deighton J., Wolpert M., Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. BMJ Open. 2016;6(8).
4. Bull C., Byrnes J., Hettiarachchi R., Downes M. A systematic review of the validity and reliability of patient-reported experience measures. Health Serv Res. 2019;54(5):1023–1035.
5. Greaves F., Jha A.K. Quality and the curate's egg. BMJ Qual Saf. 2014;23(7):525–527.
6. Black N., Varaganum M., Hutchings A. Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery. BMJ Qual Saf. 2014;23(7):534–542.
7. Doyle C., Lennox L., Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open. 2013;3(1).
8. Rivara M.B., Edwards T., Patrick D., Anderson L., Himmelfarb J., Mehrotra R. Development and content validity of a patient-reported experience measure for home dialysis. Clin J Am Soc Nephrol. 2021;16(4):588–598.
9. Zinckernagel L., Schneekloth N., Zwisler A.-D.O., Ersbøll A.K., Rod M.H., Jensen P.D., et al. How to measure experiences of healthcare quality in Denmark among patients with heart disease? The development and psychometric evaluation of a patient-reported instrument. BMJ Open. 2017;7(10).
10. Taylor R.M., Fern L.A., Solanki A., Hooker L., Carluccio A., Pye J., et al. Development and validation of the BRIGHTLIGHT Survey, a patient-reported experience measure for young people with cancer. Health Qual Life Outcomes. 2015;13(1).
11. Bosworth A., Cox M., O'Brien A., Jones P., Sargeant I., Elliott A., et al. Development and validation of a patient reported experience measure (PREM) for patients with rheumatoid arthritis (RA) and other rheumatic conditions. Curr Rheumatol Rev. 2015;11(1):1–7.
12. Bobrovitz N., Santana M., Kline T., Kortbeek J., Stelfox H.T. Prospective cohort study protocol to evaluate the validity and reliability of the Quality of Trauma Care Patient-Reported Experience Measure (QTAC-PREM) BMC Health Serv Res. 2013;13(1):98.
13. Jenkinson C. The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries. International J Qual Health Care. 2002;14(5):353–358.
14. Bastemeijer C.M., Boosman H., Zandbelt L., Timman R., De Boer D., Hazelzet J.A. Patient experience monitor (PEM): the development of new short-form picker experience questionnaires for hospital patients with a wide range of literacy levels. Patient Related Outcome Measures. 2020;11:221–230.
15. Giordano L.A., Elliott M.N., Goldstein E., Lehrman W.G., Spencer P.A. Development, implementation, and public reporting of the HCAHPS survey. Med Care Res Rev. 2010;67(1):27–37.

16. De Rosis S., Cerasuolo D., Nuti S. Using patient-reported measures to drive change in healthcare: the experience of the digital, continuous and systematic PREMs observatory in Italy. BMC Health Serv Res. 2020;20(1).
17. Corazza I., Gilmore K.J., Menegazzo F., Abols V. Benchmarking experience to improve paediatric healthcare: listening to the voices of families from two European Children's University Hospitals. BMC Health Serv Res. 2021;21(1).
18. Decourcy A., West E., Barron D. The National Adult Inpatient Survey conducted in the English National Health Service from 2002 to 2009: how have the data been used and what do we know as a result? BMC Health Serv Res. 2012;12(1):71.
19. Coulter A., Locock L., Ziebland S., Calabrese J. Collecting data on patient experience is not enough: they must be used to improve care. BMJ. 2014;348(mar26 1):g2225.
20. Davies E. Hearing the patient's voice? Factors affecting the use of patient survey data in quality improvement. Qual Saf Health Care. 2005;14(6):428–432.
21. Kunneman M., Montori V.M., Shah N.D. Measurement with a wink. BMJ Qual Saf. 2017;26(10):849–851.
22. Cunningham M., Wells M. Qualitative analysis of 6961 free-text comments from the first National Cancer Patient Experience Survey in Scotland. BMJ Open. 2017;7(6).
23. Riiskjaer E., Ammentorp J., Kofoed P.E. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. International J Qual Health Care. 2012;24(5):509–516.
24. Garcia J., Evans J., Reshaw M. "is there anything else you would like to tell us" – methodological issues in the use of free-text comments from postal surveys. Qual Quant. 2004;38(2):113–125.
25. Cammel S.A., De Vos M.S., Van Soest D., Hettne K.M., Boer F., Steyerberg E.W., et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Med Inform Decis Mak. 2020;20(1).
26. Arditi C., Walther D., Gilles I., Lesage S., Griesser A.-C., Bienvenu C., et al. Computer-assisted textual analysis of free-text comments in the Swiss Cancer Patient Experiences (SCAPE) survey. BMC Health Serv Res. 2020;20(1).
27. Van Buchem M.M., Neve O.M., Kant I.M.J., Steyerberg E.W., Boosman H., Hensen E.F. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM) BMC Med Inform Decis Mak. 2022;22(1).
28. Carlson M.L., Link M.J. Vestibular Schwannomas. N Engl J Med. 2021;384(14):1335–1348.
29. Soulier G., Van Leeuwen B.M., Putter H., Jansen J.C., Malessy M.J.A., Van Benthem P.P.G., et al. Quality of life in 807 patients with vestibular schwannoma: comparing treatment modalities. Otolaryngology–Head and Neck Surgery. 2017;157(1):92–98.
30. Gerteis M., Edgman-Levitan S., Daley J., Delbanco T.L. Jossey-Bass; San Francisco: 1993. Through the Patient's eyes: Understanding and promoting patient-centered care.
31. van Leeuwen B.M., Herruer J.M., Putter H., Jansen J.C., van der Mey A.G., Kaptein A.A. Validating the Penn Acoustic Neuroma Quality Of Life Scale in a sample of Dutch patients recently diagnosed with vestibular schwannoma. Otol Neurotol. 2013;34(5):952–957.

32. Shaffer B.T., Cohen M.S., Bigelow D.C., Ruckenstein M.J. Validation of a disease-specific quality-of-life instrument for acoustic neuroma. Laryngoscope. 2010;120(8):1646–1654.
33. Statistics Netherlands. 2017. Standaard Onderwijsindeling 2016, Den Haag.
34. Kanzaki J., Tos M., Sanna M., Moffat D.A., Monsell E.M., Berliner K.I. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. Otol Neurotol. 2003;24(4):642–648.discussion 648–9.
35. Braithwaite J., Wears R.L., Hollnagel E. Resilient health care: turning patient safety on its head. International J Qual Health Care. 2015;27(5):418–420.
36. Bastemeijer C.M., Boosman H., Van Ewijk H., De Jong-Verweij L.M., Voogt L., Hazelzet J. Patient experiences: a systematic review of quality improvement interventions in a hospital setting. Patient Related Outcome Measures. 2019;10:157–169.
37. Golda N., Beeson S., Kohli N., Merrill B. Analysis of the patient experience measure. J Am Acad Dermatol. 2018;78(4):645–651.
38. Wong E., Mavondo F., Horvat L., McKinlay L., Fisher J. Victorian healthcare experience survey 2016–2018; evaluation of interventions to improve the patient experience. BMC Health Serv Res. 2021;21(1).
39. Kleiss I.I., Kortlever J.T., Karyampudi P., Ring D., Brown L.E., Reichel L.M., et al. A comparison of 4 single-question measures of patient satisfaction. J Clin Outcomes Manag. 2020;27.
40. Salman A.A., Kopp B.J., Thomas J.E., Ring D., Fatehi A. What are the priming and ceiling effects of one experience measure on another? J Patient Exp. 2020;7(6):1755–1759.
41. Campbell J.L. Age, gender, socioeconomic, and ethnic differences in patients' assessments of primary health care. Qual Health Care. 2001;10(2):90–95.
42. Lyratzopoulos G., Elliott M., Barbiere J.M., Henderson A., Staetsky L., Paddison C., et al. Understanding ethnic and other socio-demographic differences in patient experience of primary care: evidence from the English General Practice Patient Survey. BMJ Qual Saf. 2012;21(1):21–29.
43. Johnson T.P. Response rates and nonresponse errors in surveys. JAMA. 2012;307(17):1805.
44. Poole K.G. Patient-experience data and Bias — what ratings Don't tell us. N Engl J Med. 2019;380(9):801–803.
45. Ahmed F., Burt J., Roland M. Measuring patient experience: concepts and methods. The Patient - Patient-Centered Outcomes Research. 2014;7(3):235–241.

Chapter 8

Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study

Marieke M. van Buchem, Ilse M. J. Kant, Liza King, Jacqueline Kazmaier, Ewout W. Steyerberg, Martijn P. Bauer

JMIR AI, 23 September 2024

8.1 Abstract

Background

Physicians spend approximately half of their time on administrative tasks, which is one of the leading causes of physician burnout and decreased work satisfaction. The implementation of Natural Language Processing-assisted clinical documentation tools may provide a solution.

Objective

This study investigates the impact of a commercially available Dutch digital scribe system, on clinical documentation efficiency and quality.

Methods

Medical students with experience in clinical practice and documentation (n=22) created a total of 430 summaries of mock consultations and recorded the time they spent on this task. The consultations were summarized using 3 methods: manual summaries, fully automated summaries, and automated summaries with manual editing. We then randomly reassigned the summaries and evaluated their quality using a modified version of the Physician Documentation Quality Instrument (PDQI-9). We compared the differences between the 3 methods in descriptive statistics, quantitative text metrics (word count and lexical diversity), the PDQI-9, Recall-Oriented Understudy for Gisting Evaluation scores, and BERTScore.

Results

The median time for manual summarization was 202 seconds against 186 seconds for editing an automatic summary. Without editing, the automatic summaries attained a poorer PDQI-9 score than manual summaries (median PDQI-9 score 25 vs 31, P<.001, ANOVA test). Automatic summaries were found to have higher word counts but lower lexical diversity than manual summaries (P<.001, independent t test). The study revealed variable impacts on PDQI-9 scores and summarization time across individuals. Generally, students viewed the digital scribe system as a potentially useful tool, noting its ease of use and time-saving

potential, though some criticized the summaries for their greater length and rigid structure.

Conclusion

This study highlights the potential of digital scribes in improving clinical documentation processes by offering a first summary draft for physicians to edit, thereby reducing documentation time without compromising the quality of patient records. Furthermore, digital scribes may be more beneficial to some physicians than to others and could play a role in improving the reusability of clinical documentation. Future studies should focus on the impact and quality of such a system when used by physicians in clinical practice.

8.2 Introduction

In recent years, the issue of burnout among physicians has been increasingly recognized within the health care sector. A survey conducted in 2017 involving 5000 physicians in the United States found that 44% exhibited at least 1 sign of burnout[1]. In response to this issue, the National Academy of Medicine established a committee dedicated to enhancing patient care through the promotion of physician well-being. The committee produced a detailed report titled Taking Action Against Clinician Burnout, which outlines the causes of burnout among physicians. A significant cause identified is the growing administrative workload[2]. The introduction of the electronic health record (EHR) has led to physicians spending up to half of their working hours on administrative duties[3-5]. Such tasks have been shown to lower job satisfaction for physicians[6] and negatively impact the physician-patient relationship[7]. Additionally, research linking the use of EHR to burnout indicates that physicians spending more time on EHR, particularly outside of regular hours, face a greater risk of experiencing burnout[8,9].

Recent advances in natural language processing (NLP) have created the possibility of automating some of these administrative tasks. One of these promises is the creation of the so-called “digital scribe.” Such a system, first described in

2018, automatically records, transcribes, and summarizes the clinical encounter[10,11]. A scoping review from 2022 presented an overview of the capabilities of digital scribes at that point in time, and showed that none of these systems had the full capability of a digital scribe[12]. The introduction of large language models has disrupted this field, with many papers describing their potential value in clinical note generation and multiple companies now offering digital scribe systems[13-15]. However, an evaluation on the potential impact of such a system on documentation time, including the assessment of quality and user experiences is not available to date. A thorough, prospective investigation of digital scribe performance and impact on routine practice is necessary to ensure the safety and effectiveness of the system. The aim of the current study is to assess the potential impact on the time spent and quality of medical summaries using a Dutch, commercially available digital scribe system.

8.3 Methods

Data

Our data set consisted of 27 recordings of mock consultations between physicians and nonmedical individuals. The consultations were structured around 26 vignettes, created by an internist. These vignettes delineated a set of symptoms, with a focus on various presentations of chest pain. Nonmedical individuals, assuming the role of patients, were provided with these vignettes. They were encouraged to develop and present a narrative surrounding the described symptoms. The participating physicians, all specialists in internal medicine from the Leiden University Medical Center, engaged with these simulated patients, applying their expertise to the scenarios presented. The average duration of the consultations was 293 (IQR 189-398) seconds.

Participants

In total, 21 medical students with experience in clinical practice and clinical documentation from Leiden University Medical Center consented to participate in our study. All students had a bachelor's degree in medicine and completed

a course in clinical documentation. The students received a compensation of €100 (US $111) for their participation

Autoscriber

Autoscriber is a web-based software application that transcribes and summarizes medical conversations (currently with support for Dutch, English, and German). The pipeline uses a transformer-based speech-to-text model, fine-tuned on proprietary clinical data for transcription and a mixture of large language models such as GPT-3.5 and GPT-4, combined with a tailored prompt structure and additional rules for summarization. The tool also has self-learning functionality, which was not evaluated in this study for practical reasons.

Summarization

All students summarized 4 consultations manually, then 8 consultations using Autoscriber, and finally 4 consultations manually to minimize a learning effect (see Figure 1). In total, students summarized 16 unique consultations.

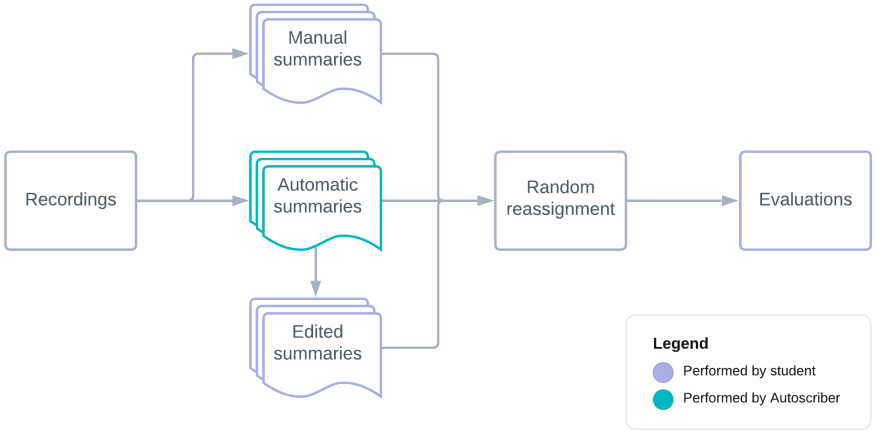


Figure 1: Flowchart showing the three different summarization methods and consecutive evaluation.

Manual summarization

Students were asked to listen to the full recording, making some notes using pen and paper. At the end of the recording, they started timing and summarized the consultation on the computer. When finished, they recorded the total time spent summarizing.

Automatic summarization

For the 8 consultations summarized using Autoscriber, the setup was similar. However, students first opened the Autoscriber application and, while listening to the recording, also recorded the consultation with Autoscriber. Once Autoscriber had created an automatic summary, students started timing and edited the automatic summary. Finally, they uploaded both the automatic summary and the edited summary, including the total time they spent editing.

Evaluation

Once all summaries were created, the manual, automatic, and edited summaries were randomly reassigned to other students, who were blinded for the method used to create the summary. Students first listened to the full recording, and then evaluated the related summaries using a modified version of the Physician Documentation Quality Instrument (PDQI-9)[16]. The PDQI-9 is a validated evaluation instrument for assessing the quality of clinical documentation, consisting of 9 questions. We removed question 1 (up-to-date: the note contains the most recent test results and recommendations) and 8 (synthesized: the note reflects the author's understanding of the patient's status and ability to develop a plan of care) for our study, as these could not be answered in the current setup. We translated the questions into Dutch, which were reviewed by one clinician (MB). Per recording, we selected the manual summary with the highest PDQI-9 score as the reference standard summary.

At the end of the study, we asked students about their experience with Autoscriber, what was positive, what should be improved, and if they would want to use Autoscriber in their work. For a more in-depth view of the differences between the automatic and edited summaries, we prompted ChatGPT (paid version, GPT-4) to assess the differences. The prompt was created iteratively

using PromptPerfect until the format of the answer was satisfactory. We then ran the prompt several times to check for internal consistency. Two researchers (MB and MvB) manually checked the answers provided by ChatGPT.

Data analysis

Preprocessing

For every summary, we calculated the total word count and the lexical diversity. Furthermore, to compare the automatic summaries to their edited counterparts we calculated the number of insertions, deletions, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE)–1 and ROUGE-L score[17], and the BERTScore metric[18]. The ROUGE-1 score calculates the overlap in words between 2 texts. The ROUGE-L score calculates the overlap based on the longest common subsequence. The BERTScore metric uses contextual embeddings to compare words between 2 texts.

Power analysis

To ensure the study was adequately powered to detect a large effect size (Cohen *f*=0.4) between 3 groups with an alpha level of 0.05 and a power of 95%, a power analysis was conducted using the FTestAnovaPower function from the statsmodels library in Python. This analysis assumed equal group sizes and did not account for potential correlations among repeated measures.

Statistical analysis

The differences between the automatic and associated edited summaries were tested using a paired t test. To compare the differences in summaries per recording, we selected the manual summary with the highest PDQI-9 score as the reference standard. We then calculated the ROUGE-1 and ROUGE-L scores for all the other manual, automatic, and edited summaries. The differences in word count, lexical diversity, PDQI-9 score, and ROUGE scores between the 3 methods was tested using one-way ANOVA and, if the P-value was below .05, followed by Tukey Honestly Significant Difference test. To assess the possibility of a learning effect, we compared the first and second batch of manual summaries on time

spent creating the summary and PDQI-9 score using a paired sample t test. We used Python for the analysis, using the "statsmodels" and the "scipy" package.

Ethical considerations

This study was conducted in accordance with the Declaration of Helsinki. For the purposes of this study, ethics approval was not applicable as the research did not include actual patients or any personal or sensitive information. All students involved in the study were informed about the purpose of the research, the use of the data, and gave their informed consent to participate in the study under these conditions.

8.4 Results

The power analysis indicated that a sample size of approximately 100 participants per group would be necessary to achieve the desired power of 95% for detecting a large effect size among the 3 groups under the specified conditions. In total, we collected 156 manual summaries, 137 automatic summaries, and 137 edited summaries from 21 students. A difference in the total number of manual, automatic, and edited summaries occurred because 3 students dropped out of the study due to time restraints. Table 1 shows an example of a manual, automatic, and edited summary of the same recording. 18 students completed the evaluation phase of the study. The median time students spent creating or editing the summaries was 186 seconds (IQR 109-267). Summaries had a median length of 129 (IQR 91-172) words. On average, summaries had a median PDQI-9 score of 28.5 (IQR 25-32) out of a maximum of 35 (Table 2). Multimedia Appendix 1 shows an extended version of Table 2, including the results of the Tukey Honestly Significant Difference test. There was a difference in time spent on manually summarizing the first batch and the second batch, with a median of 246 (IQR 137-311) and 188 (IQR 118-226), respectively, (P=.004). However, there was no difference in PDQI-9 score between these 2 batches.

Table 1: An example of a manual, automatic, and edited summary of the same recording.

Manual summary (translated)	Automatic summary	Edited summary
For a few days now, tearing pain in the chest with radiation to behind the shoulder blades. Was sitting on the couch watching TV, pain started acutely, felt like a tear and is continuously present. No complaints of tingling, numbness, or cold hands. No loss of function in arm or leg. No palpitations. Not nauseous. Has not had similar complaints before. The patient is on the "edge" of hypercholesterolemia, does not use medication. Measures blood pressure regularly for donation, pressure regular, last 120/75. The patient has never smoked. Drinks alcohol on the weekend. Family: Father had congenital vascular problems, began having heart problems at the beginning of his 50s, also had heart valve problems, died of heart failure. Brother had a heart attack at age 46, was stented. Both brothers have hypercholesterolemia. No connective tissue disorders in the family. The patient is worried due to familial CVD (Cardiovascular Disease).	Reason for referral: Main complaint: Persistent chest pain, feels as if something is tearing and radiates to behind the shoulder blades. History: Chest pain started a few days ago while calmly watching TV. The pain remains constantly present since the beginning. The pain feels as if something is tearing and radiates to the back of the shoulder blades. There are heart problems in the family (patient's father and brother). The patient's father had a poor vascular system and a leaking heart valve, and died of a heart failure. The patient's brother had a heart attack at the age of 46 and was stented. Social history: The patient does not smoke and drinks moderately alcohol on the weekend. Current medication: The patient does not use any medication. Physical examination: Additional investigation: Blood pressure is regularly measured and is about 75 over 120. The patient gives blood every six months. Diagnosis: Treatment plan: Requested examinations: Informed consent:	For several days now, tearing pain in the chest, radiating to between the shoulder blades. Chest pain started a few days ago while calmly watching TV and has been stable since then. No palpitations, not nauseous, no tingling or numb feeling. Blood pressure is regularly measured and is about 75 over 120. Cholesterol is good. Patient is worried because of family history. Fam: father had heart problems, brother had a heart attack at a young age, hypercholesterolemia, no connective tissue diseases. Intox: no smoking, alcohol on weekends in moderation Med: none

Table 2: Descriptive statistics of the different methods and associated p-values.

Metrics	Manual (n=156), median (IQR)	AS edited (n=137), median (IQR)	AS (n=137), median (IQR)	*P* value (ANOVA)
Time spent on summary (seconds)	202 (128-286)	152 (93-244)	0 (0-0)	<.001
Word count	101 (67-141)	137 (96-194)	148 (116-180)	<.001
Lexical diversity	0.68 (0.63-0.74)	0.61 (0.56-0.66)	0.59 (0.53-0.63)	<.001
PDQI-9[a] score				
Overall	31 (27-33)	29 (26-33)	25 (22-28)	<.001
Accurate	5 (4-5)	5 (4-5)	4 (2-5)	<.001
Thorough	4 (4-5)	4 (4-5)	3 (2-4)	<.001
Useful	5 (4-5)	4 (4-5)	4 (3-4)	<.001
Organized	4 (3-5)	4 (3-5)	4 (3-4)	.01
Comprehensible	5 (4-5)	5 (4-5)	4 (3-5)	<.001
Succinct	5 (4-5)	4 (2-5)	3 (2-4)	<.001
Internally consistent	5 (4-5)	5 (4-5)	5 (4-5)	<.001
ROUGE[b,c]-1 F_1-score	47.3 (42.5-56.4)	40.6 (35.0-45.4)	32.3 (27.0-37.4)	<.001
ROUGE-L F_1-score	29.4 (23.7-37.6)	23.4 (20.6-27.5)	19.6 (15.7-23.5)	<.001
BERTScore[c] F_1-score	74.6 (71.9-77.0)	71.6 (69.5-73.7)	68.6 (67.5-70.3)	<.001

[a]PDQI-9: Physician Documentation Quality Instrument.

[b]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[c]To calculate the ROUGE score and BERTScore, the highest scoring manual summary was taken as the reference standard. These summaries were taken out of the data set when calculating the average ROUGE scores.

Comparison between automatic and corresponding edited summaries

Students inserted a median of 45 (IQR 27-82) words and deleted 46 (IQR 27-80) words. The edits led to a median increase in PDQI-9 score of 4.0 (IQR 1-8). The median ROUGE-1 F1 score between the automatic and their corresponding edited summaries was 73.3 (IQR 61.0-84.4), the ROUGE-L F1 score was 67.4 (IQR 50.0-80.5), and the BERTScore F1 was 84.1 (IQR 79.0-89.4).

ChatGPT assessed the differences between automatic summaries and their edited counterparts on the following aspects: language use and precision, clarity and detail, coherence and flow, structural differences, stylistic variations, and the most common deletions and insertions. The final prompt can be seen in Multimedia Appendix 2. See Table 3 for the observations per aspect. The assessment by ChatGPT aligned with the sample analysis performed by the researchers. Furthermore, similar aspects were mentioned by the students.

Table 3: Differences between automatic and edited summaries, as assessed by ChatGPT.

Aspect	Automatic summaries	Edited summaries	Observations
Language use and precision	Generally simplistic and formulaic language. For example, "Chest pain started a few days ago while quietly watching TV".	More sophisticated and precise language. Example: "Since a few days tearing chest pain radiating to between the shoulder blades".	Human editors refine the language to be more precise and contextually appropriate.
Clarity and detail	Often vague, lacking specific details. For instance, "Patient has had persistent watery diarrhea since one week".	Provide clearer, more detailed descriptions. Example: "Patient has had persistent watery diarrhea for a week with a frequency of ten times a day".	Human editing enhances clarity by adding relevant details that were omitted in the automatic summaries.
Coherence and flow	Sometimes disjointed or lacking in logical flow. Example: "The chest pain started suddenly and has been continuously present since it started".	Better structured, with a smoother flow of ideas. Example: "The patient complains of sudden and persistent chest pain that started several days ago".	Human editors improve the coherence, making the summaries easier to follow.
Structural differences	Tend to follow a predictable structure, possibly template-based.	More varied structures, adapted to the content's needs.	Human editing allows for more flexible structuring, tailored to the specific summary.
Stylistic variations	Limited stylistic variations, often repetitive.	Display a wider range of styles, adapting to the tone and context.	Human editors introduce stylistic diversity, making each summary more unique.
Most common deletions			Redundant phrases, overly general statements.
Most common insertions			Specific details, clarifying phrases, and contextual information.

Differences per student

Using Autoscriber had a different effect per student. For 8 out of 18 students, using Autoscriber was associated with a decrease in PDQI-9 score, while for the other students the difference in PDQI-9 score between manual and automatic summaries had a P value above .05. For 5 students, editing the automatic summary took more time than manually creating a summary, although these differences were not significant. For 3 students, editing the automatic summary led to a decrease in time spent on summarizing, with a P value lower than .05. See Multimedia Appendix 3 for the full overview.

Experiences with using Autoscriber

Students were generally very positive about using Autoscriber, mentioning that it was nice or interesting to use (n=9), easy and simple in use (n=6), and that they believed in the potential of such a tool (n=4). Four students mentioned the automatic summary exceeded their expectations, while 4 other students said the quality of the summary was insufficient due to errors and the amount of time needed to make edits. A specific error that was mentioned multiple times was that the summary did not include negative symptoms (eg, the absence of shortness of breath). Three students mentioned the tool did not always work: it would sometimes load for a very long time or get stuck while generating the summary. This was due to limitations in graphics processing unit capacity at that time. See Table 4 for the positive aspects and points of improvement mentioned by the students. A majority of students (12/18, 67%) would want to use the application during their work. The other students (6/18, 33%) said they would want to use the application if improvements were made.

Table 4: Themes most often described by students about the positive aspects and points of improvement.

	Mentioned aspects	Count
Positive	Easy to use	5
	Good accuracy. E.g. amount of details, good use of language, low amount of errors, inclusion of important symptoms	5
	Summary fairly complete	4
	Saves time	4
	Well-structured view	4
	Nice to have something to start with, without typing	3
Negative	Structure does not align with preferences. E.g. headings unclear, illogical structure, does not align with style	6
	Wordy/lengthy	5
	Relevant information missing. E.g. details, absence of symptoms	5
	Comments on language use. E.g. use of non-standard words, vague descriptions, too literal, absence of common abbreviations	5
	Duration of summarization time	3
	Presence irrelevant information	2

8.5 Discussion

In this impact study, we extensively evaluated the efficacy of Autoscriber, a Dutch digital scribe system, in enhancing the clinical documentation process in a pilot setting. A group of trained medical students summarized clinical conversations with and without the tool. We found differences between automatic and manual summaries in time spent on the summary, the word count, lexical diversity, and qualitative aspects such as accurateness and usefulness. These differences decreased after students edited the automatic summaries. During editing, medical students most often added context and details, while removing overly general statements and irrelevant text. Most were positive about using the tool, although some mentioned the summaries were lengthy and the structure did not always align with their preferences.

As the first impact study of a fully functioning digital scribe system, we provide some interesting insights into the possible future of digital scribes in health care. First of all, we show that a collaboration between the system and the students

leads to the best results at this point in time, with a decrease in time spent on summarizing in combination with a similar quality when compared to manual summarization. We believe the current setting might even provide an overestimation of the quality of the manual summaries: the students did not have a time cap for creating the summaries, while in clinical practice, physicians often have to create a summary during or in between consultations. Furthermore, multiple studies show a negative association between seniority of a physician and the completeness of a medical record[19-21]. Taking this into account, we see the potential in using a digital scribe system that provides a first draft, which the physician then edits. In the current setup, this collaboration led to a decrease in time spent summarizing, while keeping the quality of the summary on par.

When looking at the differences between the 3 methods, the higher word count and lower lexical diversity in the automatic summaries compared to the manual summaries stood out. Two previous studies compared human and ChatGPT-written medical texts and reported similar results[22,23]. Furthermore, one of these studies reported human texts contained more specific content, which we found as well. These aspects are essential to improve in future versions, as they directly link to the quality of a summary in terms of succinctness and thoroughness. An increased summary length could lead to an increase in time spent reading or analyzing summaries downstream in the clinical process. However, a small decrease in lexical diversity in combination with a more structured summary could also be seen as a step toward standardization of medical summaries. This aspect is becoming more important since clinical documentation is increasingly reused for other purposes, such as research and quality measurements. Furthermore, previous studies show that structured documentation leads to increased note quality[24], which in turn has been shown to positively affect the quality of care[25-27]. These potential effects have to be studied in future research.

We found large differences in the effect of using Autoscriber on PDQI-9 score and time spent summarizing between students. While using Autoscriber decreased the time spent on finalizing the summary for most students, there were a few students who spent more time on editing the automatic summary then on manually creating a summary. Furthermore, the difference in PDQI-9 score

between manual and automatic summaries differed greatly between students. This result is highly relevant, as it shows that the added value of using a digital scribe differs per user. Future studies should investigate which users could gain most benefit in using a digital scribe, taking into account age, specialty, the ability to type blindly, and other factors that might impact the added value on a personal level.

Strengths and limitations

This impact study on a digital scribe system for clinical conversations presents a novel exploration into the practical application of such technology. Since the introduction of ChatGPT, many papers have described the potential of using ChatGPT and other large language models in health care. While their potential is clear, these models have still to prove their actual clinical value. This study takes a first step in gaining a better view of the potential effects such a digital scribe system could have on the documentation process, especially in interaction with the user. Apart from quantitative analyses, we also included several different qualitative analyses, providing a more in-depth view of the differences between the summaries and the experiences of the students. These results are highly relevant for researchers and companies developing digital scribes as well as health care organizations considering using a digital scribe in the near future.

One limitation is the setup of our study, which is not fully representative of clinical practice. Specifically, our reliance on medical students listening to pre-recorded mock consultations does not fully capture the dynamic and often unpredictable nature of real-time clinical interactions. The controlled environment of our study does not account for the varied technological, environmental, and personal factors that can influence the use and effectiveness of digital scribe systems in live clinical environments. However, this approach allowed us to isolate and evaluate the impact on summarization time and differences in summary between the 3 methods. Future research should aim to incorporate real clinical interactions to validate and extend our findings.

Another limitation is the lack of a reference summary per consultation. To calculate the ROUGE scores, we designated the highest scoring manual summary

as the reference standard per consultation. This method suffices for the current pilot study; however, it brings up the bigger issue of summary evaluation metrics. The ROUGE score remains the most used metric, while this metric only measures exact overlap in words and is, thus, very sensitive to the choice of reference summaries[28]. Because of this limitation, we added the BERTScore metric, which has been shown to correlate better with human evaluations[18]. However, the overall lack of a standard for clinical documentation still poses a considerable challenge for the objective assessment of summarization efficacy of digital scribes. This underscores the necessity for developing more sophisticated evaluation methods, especially with the arrival of large language models in health care.

Future implications

Our findings underscore the promising potential of integrating digital scribe technologies like Autoscriber within clinical settings to alleviate the administrative burdens faced by health care professionals. Future clinical impact studies are imperative to explore the broader effects of digital scribes on the physician-patient interaction, documentation accuracy, and overall health care delivery efficiency. These studies should aim to evaluate the real-world applicability of digital scribes, including their impact on clinical workflow, quality of care, and patient satisfaction. Especially the latter, which has not received sufficient attention up to now, should be the focus of future research to ensure the physician-patient relationship is not harmed. Additionally, exploring the customization of digital scribe systems to fit the specific needs and preferences of individual physicians or specialties could enhance user adoption and effectiveness. As the field of large language models is developing at a fast rate and digital scribes will improve quickly, repeated or continuous evaluation of these systems is necessary. A recent study described the development and evaluation of a chat-based diagnostic conversational agent[29]. This agent outperformed primary health care providers in both diagnosis and the development of a treatment plan. The introduction of digital scribes in clinical practice could eventually lead to similar support during the clinical encounter, where the digital scribe might suggest additional follow-up questions or provide a differential diagnosis. Ultimately,

the goal is to seamlessly integrate digital scribes into clinical practice, ensuring they enhance patient care and physician well-being.

Conclusion

This study explores the impact of a Dutch digital scribe system on the clinical documentation process, offering significant insights into its potential to enhance physicians' experience. By demonstrating the use of the system in reducing summarization time while maintaining summary quality through collaborative editing, our research highlights the potential of digital scribe systems in addressing the challenges of clinical documentation. Despite the limitations related to the representativeness of our pilot setup and the evaluation of summary quality, the positive outcomes suggest a promising avenue for future research and development. Further studies, particularly those involving real-world clinical settings, are essential to fully understand the implications of digital scribes on the physician-patient dynamic and health care delivery.

Data availability

The majority of the data supporting the findings of this study are included in the manuscript. Additional data sets generated during and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request and subject to approval by the institutional review board.

Conflicts of interest

JK, LK, and MB are employees of Autoscriber. Their affiliation with Autoscriber did not influence the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The other authors, who are not affiliated with Autoscriber, contributed independently to this work, ensuring unbiased data interpretation and conclusions.

Supplemental information

Multimedia Appendix 1

Multimedia Appendix 2

Multimedia Appendix 3

References

1. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, Carlasare LE, Dyrbye LN. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681–1694. PMID:30803733
2. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. The National Academies Press; 2019.
3. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W-J, Sinsky CA, Gilchrist VJ. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Medicine 2017;15(5):419–426. PMID:28893811
4. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, Westbrook J, Tutty M, Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016;165(11):753. PMID:27595430
5. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, Wang W, Luft HS. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. Health Affair 2017;36(4):655–662. PMID:28373331
6. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, Torchiana DF. The Impact of Administrative Burden on Academic Physicians. Acad Med 2017;92(2):237–243. PMID:28121687
7. Pelland KD, Baier RR, Gardner RL. "It's like texting at the dinner table": A qualitative analysis of the impact of electronic health records on patient-physician interaction in hospitals. J Innovation Heal Informatics 2017;24(2):216–223. PMID:28749316
8. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, Linzer M. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assn 2019;26(2):106–114. PMID:30517663
9. Robertson SL, Robinson MD, Reid A. Electronic Health Record Effects on Work-Life Balance and Burnout Within the I3 Population Collaborative. J Graduate Medical Educ 2017;9(4):479–484. PMID:28824762
10. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. Npj Digital Medicine 2018;1(1):58. PMID:31304337
11. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digital Medicine 2019;2(1):114. PMID:31799422
12. Buchem MM van, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. Npj Digital Medicine 2021;4(1):57. PMID:33772070
13. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–1940. PMID:37460753
14. Giorgi J, Toma A, Xie R, Chen SS, An KR, Zheng GX, Wang B. WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models. arXiv 2023; doi: 10.48550/arxiv.2305.02220

15. Abacha AB, Yim W, Fan Y, Lin T. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. Proc 17th Conf Eur Chapter Assoc Comput Linguistics 2023;2291–2302. doi: 10.18653/v1/2023.eacl-main.168

16. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;03(02):164–174. PMID:22577483

17. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out Barcelona, Spain: Association for Computational Linguistics; p. 74–81.

18. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv:1904.09675. 2019.

19. Lee FCY, Chong WF, Chong P, Ooi SBS. The emergency medicine department system: a study of the effects of computerization on the quality of medical records. Eur J Emerg Med 2001;8(2):107–115. PMID:11436906

20. Lai FW, Kant JA, Dombagolla MH, Hendarto A, Ugoni A, Taylor DM. Variables associated with completeness of medical record documentation in the emergency department. Emerg Med Australas 2019;31(4):632–638. PMID:30690885

21. Soto CM, Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. BMC Heal Serv Res 2002;2(1):22. PMID:12473161

22. Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue J, Wu Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv 2023; doi: 10.48550/arxiv.2301.07597

23. Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y, Huang X, Zhu D, Cai H, Li Q, Liu T, Li X. Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study. JMIR Méd Educ 2023;9:e48904. PMID:38153785

24. Ebbers T, Kool RB, Smeele LE, Dirven R, Besten CA den, Karssemakers LHE, Verhoeven T, Herruer JM, Broek GB van den, Takes RP. The Impact of Structured and Standardized Documentation on Documentation Quality; a Multicenter, Retrospective Study. J Méd Syst 2022;46(7):46. PMID:35618978

25. Elkbuli A, Godelman S, Miller A, Boneva D, Bernal E, Hai S, McKenney M. Improved clinical documentation leads to superior reportable outcomes: An accurate representation of patient's clinical status. Int J Surg 2018;53:288–291. PMID:29653245

26. Reyes C, Greenbaum A, Porto C, Russell JC. Implementation of a Clinical Documentation Improvement Curriculum Improves Quality Metrics and Hospital Charges in an Academic Surgery Department. J Am Coll Surg 2017;224(3):301–309. PMID:27919741

27. Kittinger BJ, Matejicka A, Mahabir RC. Surgical Precision in Clinical Documentation Connects Patient Safety, Quality of Care, and Reimbursement. Perspect Heal Inf Manag 2016;13:1f. PMID:26903784

28. Akter M, Bansal N, Karmaker SK. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? Find Assoc Comput Linguistics: ACL 2022 2022;1547–1560. doi: 10.18653/v1/2022.findings-acl.122

29. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S, Singhal K, Cheng Y, Hou L, Webson A, Kulkarni K, Mahdavi SS, Semturs C, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Karthikesalingam A, Natarajan V. Towards Conversational Diagnostic AI. arXiv 2024; doi: 10.48550/arxiv.2401.05654

Part 3

General Discussion and Summary

Chapter 9

General Discussion

The general aim of this thesis was to explore the application and potential value of natural language processing (NLP) in healthcare. More specifically, we studied the development and validation of NLP models in different settings, showing promising results but also highlighting challenges that need to be overcome for responsible implementation in clinical practice. Furthermore, we assessed the value of two NLP models in a pilot study. In this chapter, the main findings will be discussed and recommendations for further research will be provided.

9.1 Promising applications

NLP provides many opportunities for application in healthcare. We examine which NLP tasks hold most promise for healthcare settings. We then shift our focus to the different data sources within the healthcare environment, analyzing how NLP can be applied to these datasets to improve care delivery. An overview of promising applications is presented in Table 9.1.

9.1.1 Opportunities of various NLP tasks

Classification

Classification is the most common task in clinical natural language processing[1], which is also reflected in the studies presented in this thesis. The models we developed vary in performance, which may be due to several causes, of which the amount of training data is an important one**.** To be able to generalize to new data, a model needs to have seen a representative number of examples per class during training. The classification model for distinguishing negative from non-negative patient experiences (Chapter 4) had a lower performance than the model for distinguishing neutral from positive, which might be explained by the low number of examples for the negative experiences. Thus, we should be critical about the feasibility of training a classification model from scratch in settings with a low number of examples per label. However, clear guidelines on how much data is needed is lacking. Recent advances in NLP, such as the introduction of pretrained language models, have made it possible to train models with less training data[2–4]. The latest transformer models, such as GPT3 and later versions and Llama2, even allow for zero-shot and few-shot learning, where the model is provided zero training data or just a few examples[5,6].

Topic modeling

In Chapters 4 and 6 we had large, unlabeled datasets of patient experiences and social media data, that were subject to change over time. Topic modeling proved useful in both these settings: in the patient experiences setting (Chapter 4 and 7), it led to the identification of three new action points by the care team; in the statin setting (Chapter 6), the primary topics aligned with previous findings, such as low adherence due to adverse effects or preference for lifestyle alternatives. However, some topics uncovered novel points of discourse, such as the role of statins in improving COVID-19 outcomes, which could be interesting for future research. Within healthcare, multiple studies have used topic modeling to uncover patient perspectives from social media data with similar findings[7–11].

Other settings where topic modeling has proven useful are, for example, automatic phenotyping of patients using EHR data[12,13]. However, topic modeling is mostly useful for data exploration and as part of an analysis pipeline. Without any postprocessing, the topic descriptions are difficult to interpret and thus not suitable for settings where a clear label is needed. Often, the resulting topics are used as input to a classification model, using topic modeling for dimensionality reduction[12,13]. The recent introduction of large language models (LLMs) might offer a solution to the lack of interpretability by being able to describe the contents of the different topics[14]. This might improve the applicability of topic modeling on text data in a healthcare context. Furthermore, large language models might be able to take over the full topic modeling pipeline, as these models have unprecedented summarization capabilities. To date, only one study performed this comparison. This study shows that LLMs can be a robust substitute for current topic modeling techniques, although extensive prompt engineering is necessary to extract useful topics[15].

Summarization

The task of summarizing texts has seen a huge jump in performance with the introduction of LLMs. At the time of writing the scoping review (Chapter 2), most studies used extractive summarization, entity extraction, and classification to filter and sort the information relevant for a summary, although this did not lead to natural, narrative summaries. The introduction of LLMs has made

all these previous techniques obsolete for summarization purposes. Our pilot study with Autoscriber (Chapter 8) shows the high quality of these summaries: medical students are positively surprised by the quality of the summaries and a majority wants to use this technique in their work. In the past few months, several similar studies have been published, showing promising performance of LLMs in creating discharge summaries[16] and clinical information letters for patients[17,18], and summarizing radiology reports[19–21], patient-doctor dialogues[20,22–24], and the many handovers between shifts or departments happening within the hospital every day[25]. Many of these opportunities address the current administrative burden in healthcare, while potentially improving the quality of the collected data because of increased consistency, structure, and objectiveness. These opportunities are not without challenges and risks, which will be discussed in Section 9.3.1. Furthermore, the major challenge related to summarization is the difficulty in creating a reference standard. This topic will be discussed in Section 9.2.1.

9.1.2 Opportunities of various data sources

In this thesis, we focused on three data types: clinical data, patient-generated data, and social media data. Within the literature, clinical data such as EHR data is the most common data source used within the clinical NLP field[26]. As most information about the patient is captured within free text EHR data, a large focus within the clinical NLP field has been to use this data to develop patient risk classification and prediction models[27]. In our acute care utilization setting (Chapter 3) we show that models trained on only free-text data perform similarly to models trained only on structured data, without the need for extensive preprocessing. This finding is in line with other studies, underlining the value of free text data relative to structured data[28,29]. A future research question would be if the use of free text EHR data allows for easier transfer of models between organizations. There are, however, many challenges related to the use of EHR data, which we will discuss in Section 9.2.1. A new type of clinical data is the recorded and transcribed clinical conversation. With the introduction of large language models, it has become possible to leverage the value of this data type. Apart from the previously discussed summarization capabilities of these models, the application of LLMs on clinical conversational data leads to a

new level of granularity in clinical data. Having complete transcripts of clinical conversations might lead to more, in-depth and structured data about disease trajectories and patient presentations of various diseases, potentially paving the way for improved clinical decision support. No literature is available on this topic yet.

Another popular data source due to its accessibility is social media data. We investigated different ways in which social media data might improve clinical practice in our statin setting (Chapter 6) and patient messages setting (Chapter 5). In the former, social media was used to gain insights into public views on statins. The results of this study can be used to inform clinicians which topics to discuss when talking to patients about statins. In the latter, we used continued pretraining on Reddit data, aiming to improve the performance of a pretrained language model on identifying depression concerns from patient messages. In this case, this process did not improve the overall performance of the model. With respect to direct value for healthcare, social media allows for extracting patient perspectives, coping strategies, and even adverse drug events[30–33]. Furthermore, social media data is being used to pretrain (large) language models such as Llama 2[5].

A relatively underused data type is patient-generated data, such as patient experience data and patient messages. The potential benefits of this data type are underlined in our patient experience and patient messages settings (Chapters 3, 4, and 7). We see that capturing free-text patient experiences leads to new insights, which can be turned into concrete points of improvement, thus having the potential to directly improve patients' experiences. In Chapter 5, we see that patient messages show signs of distress that can be used to identify depression concerns. These settings show the value of this data type, along with the value of using NLP to take away the burden of analyzing all the incoming data from healthcare professionals. Especially patient messages are underused, while these could be very valuable in creating triggers for mental health issues or other acute care needs. Recent studies mostly highlight the burden, potential value, and characteristics of this data type, but examples in clinical practice are lacking[34–37]. Recent studies use LLMs to create patient-focused chatbots or

create draft replies to patient messages, with mixed results [38,39]. Although it might have the potential to make healthcare organizations more accessible to patients and free up time from healthcare professionals, this has not been shown in large-scale evaluations.

Table 9.1: an overview of the most promising applications of NLP in healthcare.

Data source & task	Application
Clinical data & summarization	Use NLP models with summarization capabilities to automatically generate clinical notes, discharge letters, and summaries of patient files.
Clinical data & classification	Use BERT and LLMs to train classifiers in settings with a small amount of training data. Train classification models on free-text data to simplify the transfer of models between organizations.
Patient-generated data & classification	Apply classification models to patient messages to create triggers for mental health or acute care needs.
Patient-generated data & summarization	Use NLP models with summarization capabilities to provide relevant information to patients.
Social media data & topic modeling	Use topic modeling techniques in combination with LLMs to extract general perspectives about health-related topics from social media data.

9.2 Challenges during development

In the course of developing NLP models for the settings outlined in Chapters 3 to 6, we faced multiple challenges that hindered their adoption in clinical settings. This section details these challenges, including issues related to data, bias, explainability, and deployment. See Figure 1 for an overview.

9.2.1 Data

As with other types of machine learning, NLP models are only as good as the data they are trained on. Especially in healthcare, access to large amounts of high-quality data is a huge challenge. This challenge encompasses the quality of the data itself, its availability and accessibility, and the choice of a trustworthy reference standard.

Data quality

Most clinical NLP models are trained on EHR data, which is not primarily collected for research or development purposes but to create a complete record of clinical information and proceedings[40]. The lack of standardized methods to create free-text clinical notes in combination with the wide variety in the human use of language leads to large heterogeneity of free-text clinical data between clinicians, departments, and organizations[40]. A recent study reported differences in medical coding practices between two organizations due to a different billing system, affecting clinicians' incentive to code[41]. Furthermore, diagnostic coding systems such as ICD-10 are primarily used for billing purposes and thus may not always correspond to the actual diagnosis[42]. Other studies report high rates of missing problems in the problem list, especially if physicians are overstretched[43,44]. In our acute care utilization case (Chapter 3), an extra preprocessing step was added to reduce the large number of redundant notes that were found in patients' records[45,46]. However, NLP might provide a solution to this challenge. Many studies use NLP to improve the quality and structure of EHR data, such as extracting the presence of symptoms or the medication dosage and linking terms to ontologies (e.g. UMLS[47], ICD-10[48], or SNOMED-CT[49])[50].

Data availability

An essential aspect of developing NLP tools for healthcare is having data available for training purposes. Within the healthcare context, data availability can be challenging due to the challenges around data sharing. This holds for structured data, but even more so for text data as it is not always possible to fully anonymize the data. As seen in Chapters 2 through 8, this means most studies include data from a single institution. This trend is seen across the whole clinical NLP field, with most studies using proprietary data[1,50]. In the past few years, many solutions have been proposed to tackle this challenge. Federated learning is a technique where the machine learning model is trained on data from multiple organizations, but instead of the data leaving the organizations to be collected in a central location, the model 'travels' from organization to organization[51]. Another promising solution is the development of synthetic data, where a model is trained to create new data that is very similar to the training data. Using this

technique, a hospital can create a new dataset that is shareable with other organizations, without sharing actual patient data[52,53].

Reference standard

Reference standards are essential for supervised NLP methods requiring labeled data, but also for evaluating the output of unsupervised methods. While some settings will have very clear reference standards (such as in-hospital mortality), finding a trustworthy reference standard is a challenge in many settings. Apart from reference standard issues related to data quality and data availability as discussed in the previous sections, other challenges arise when a reference standard is not present. In our patient messages setting (Chapter 5) we manually labeled a large dataset of patient messages as concerning or not concerning. We defined 'concerning' with a group of clinicians and created an extensive annotation guideline. Even so, the inter-annotator agreement was 0.38, which is considered moderate. Such a score is not uncommon in the healthcare domain. A study describing the development of a large corpus of annotated clinical conversations reports similar scores, especially for entities such as symptoms and conditions[54]. The topic of inter-annotator agreement or inter-rater reliability has been described repeatedly in medical literature[55,56]. However, it takes on a different meaning when these labels are not just used for quality control, but for training AI models that might take over tasks from trained clinicians. Although in some cases a single truth can be unveiled, in many cases trained clinicians will have irreconcilable disagreements about ground truth labels. A growing body of research recognizes the importance of including these differing opinions in the development process to improve the generalizability of the model and, most importantly, decrease bias[57,58]. Furthermore, a recent paper argues that machine learning in healthcare should strive to move past the quality and accuracy of current clinical practice[59]. They recommend using objective ground truth metrics such as mortality or patient centered metrics such as perceived pain and, where possible, taking ambitious steps to elevate existing standards.

9.2.2 Bias

With the current large-scale development of NLP models in healthcare, we must pay attention to who is benefiting from these models and who is potentially

harmed. Especially when using patient-generated data, we must be conscious of the differences in access to certain communication methods and patients' ability to express themselves. In the evaluation of our patient experience model (Chapter 7), we found that patients who filled out the questionnaire had a higher level of education. Furthermore, of the patients that completed the questionnaire, patients who were male and who had a low level of education more often responded with only one word. In our patient messages setting (Chapter 5), our cohort consisted of mostly White and Asian, privately insured patients. In this last example, we risk perpetuating existing biases in the (lack of) recognition of depression in different populations. Both examples show the need to critically reflect on who has access to the tools we develop[60–62].

A more recent problem in relation to bias is the data that (large) language models are trained on. Our own RedditBERT model (Chapter 5) was trained on Reddit data, which is not representative for patients in general. Autoscriber (Chapter 8) uses GPT3.5, which has a bias towards dominant values from the United States because of the available data for training[63]. A recent paper reviewed the current clinical language models and found that almost all of them were trained on the MIMIC-III dataset[64]. This dataset contains data from the intensive care unit of the Beth Israel Deaconess Medical Center, an academic medical center in Boston, which will not be representative for large parts of healthcare[59]. There are many ways to mitigate bias, including guidance ethics, increasing diversity of AI developers, and increasing diversity within the data. Guidance ethics is a method developed in The Netherlands that uses a multistakeholder perspective to uncover the potential effects of the model, aiming to strengthen the positive effects while mitigating the negative effects[65]. An essential part of guidance ethics is including a representative and diverse group of people to include different perspectives, which is also important for the AI community as a whole. We discuss this topic in our commentary 'Picture a data scientist', providing several recommendations[66]. Lastly, several initiatives are working on creating datasets that include a diverse representation of the population. The All of Us research program is a great example, with more than 80% of its participants belonging to groups that have historically been underrepresented in biomedical research[67].

9.2.3 Clinically relevant metrics

Many papers presenting NLP models do not publish performance metrics that are clinically relevant. Within the machine learning community, it is standard practice to report metrics such as the precision and recall based on the performance of the model on an unseen test set. Obviously, these statistical performance metrics do not consider how the model will be used in clinical practice and if the output of the model is clinically useful and which actions can subsequently be taken[42,59,68]. Furthermore, these metrics do not provide a comprehensive view of whether an NLP model will deliver value in healthcare settings, making it challenging to discern which models are promising for further development and which ones may not be useful. For every setting, researchers should critically assess which performance metrics are relevant and in line with clinical goals and workflows. Good examples of clinically relevant metrics include decision curve analysis[69] and the number needed to benefit[70]. A recent paper proposed an extensive evaluation framework for healthcare chatbots, including an evaluation of the interface and interaction with the user, and highlighting the importance of manual evaluations by end users[71]. Ultimately, value must be evaluated in a clinical trial, which will be discussed in Section 9.3.3. However, the metrics suggested here can provide valuable insights during development.

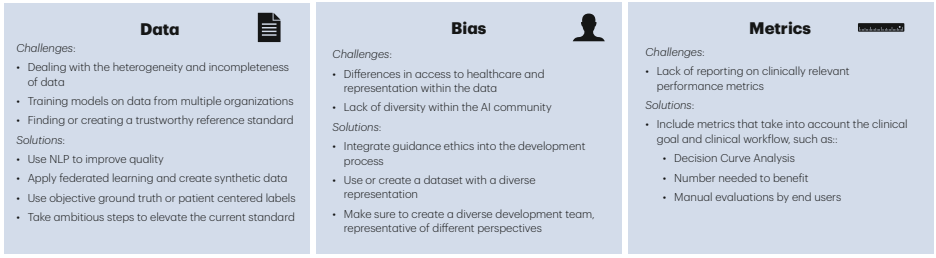


Figure 1: an overview of the challenges during development of NLP models for healthcare.

9.3 Value for clinical practice

Although the promise of applying natural language processing to healthcare is undeniable and vast amounts of NLP tools are developed, scientific reports on the value of these tools when deployed in clinical practice are lacking. Our

scoping review (Chapter 2) highlights this in the realm of digital scribes, and similar patterns are observed in other areas[72]. Among the models discussed in Chapters 2 through 6, two have been developed into applications, either by integration with existing systems or as standalone tools. This section discusses the definition of value for clinical practice, how to create value with NLP applications, and how to evaluate this in clinical practice (see Figure 2).

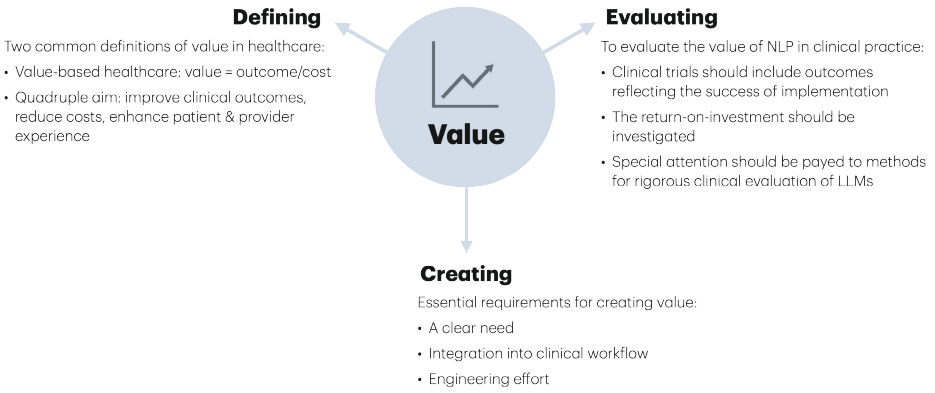


Figure 2: an overview of defining, creating, and evaluating value.

9.3.1 Defining Value in Healthcare

A first step in investigating the value of NLP for healthcare is defining what value for healthcare encompasses. The most widespread approach to value for healthcare has been the value-based healthcare (VBHC) approach introduced by Porter and Teisberg[73]. They argue that value within healthcare should be defined around the patient and can be calculated as follows:

$$Value = \frac{Outcome}{Costs}$$

These outcomes go beyond traditional metrics such as medical procedures or tests and encompass a broader spectrum of factors such as patient-reported outcomes (PROs), functional status, quality of life, and long-term health outcomes. Thus, VBHC focuses on weighing the impact of healthcare interventions

on patients' health status and overall well-being against their costs. Another common approach towards assessing value within healthcare is the Quadruple Aim[74,75]. This approach describes value in similar terms as VBHC, including patient outcomes, costs, and patient experience, but also adds an extra dimension: clinician experience (see Figure 3).

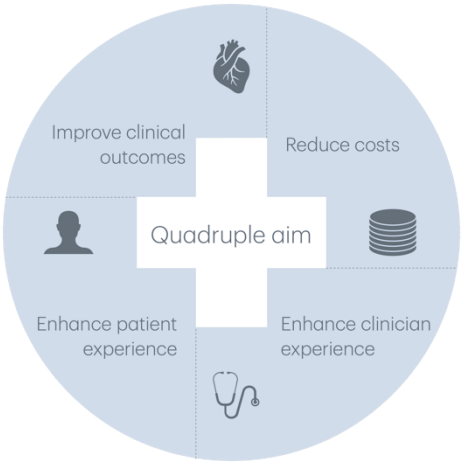


Figure 3: the four dimensions of the Quadruple Aim.

In terms of these definitions of value for healthcare, the promise of NLP lies in several dimensions. The first and most straightforward dimension is the clinician experience. With an increasing rate of clinician burnout and decreased job satisfaction, there is an urgent need for improvement[76–80]. With the promise of taking over tasks such as writing discharge letters, generating clinical notes, and replying to patient queries, NLP might have a serious impact [81]. Furthermore, several studies have shown the ability of NLP to perform diagnostic tasks[82,83]. These developments could lead to a decrease in resources needed for diagnostic work-up. It is still speculative at this point in time whether costs and patient outcomes improve, since large, high-quality clinical trials are lacking.

9.3.2 Creating Value with NLP Applications

Within this thesis, we did not reach the clinical trial stage with any of the NLP models. However, we performed two pilot studies, and together with findings

from the literature they highlight essential requirements for creating value with NLP applications: a clear need, integration into the clinical workflow, and engineering effort.

A clear need

Both the AI-PREM (Chapter 7) and Autoscriber (Chapter 8) originated from clear needs from clinical practice. As pointed out in recent research, it is crucial to distinguish between 'machine learning on healthcare data' and 'machine learning for healthcare problems'[59]. While the former advances the field theoretically, it is the latter that tends to create actual value in healthcare. Developers and researchers need to step away from primarily developing models for settings where large datasets and labels are readily available and engage with healthcare professionals to find the right problems[42]. This is a fundamental step towards realizing clinical value.

Integration into the clinical workflow

A recent commentary states that the key to achieving meaningful impact with AI is focusing on behavioral change and thus on changing routines and care processes[84]. For the AI-PREM and Autoscriber, significant effort was invested in determining how they would fit into the existing clinical workflows. This process involved extensive discussions with a multidisciplinary team over many iterations. In the past few years, a few useful guidelines have been published, describing the process from idea to implementation in healthcare[85,86]. These guidelines specifically pay attention to integrating the model into the clinical workflow.

Engineering effort

Another essential requirement for workflow integration is the engineering effort needed to develop an application or integrate into an existing application. For the AI-PREM, we chose to integrate it into an already existing dashboard, while Autoscriber is currently a standalone web application that will soon be able to integrate with the EHR. These engineering efforts demand considerable effort and expertise and are currently often neglected in development projects[59]. To achieve widespread clinical value, it is essential to simplify these engineer-

ing steps. In radiology, for example, several companies have begun offering 'AI platforms' that provide plug-and-play functionality for various AI applications, facilitating large-scale integration into hospital systems[87,88].

9.3.3 Evaluating Value

Since high-quality, large-scale clinical trials with NLP models are lacking, it is important to reflect on how to best evaluate the value of these models in clinical practice[81]. As discussed in our scoping review (Chapter 2), many studies lack error analyses, and clinical validation or utility assessments are almost non-existent. These assessments are paramount in gaining clinicians' trust in these models. A few recent perspectives provide guidance on how to evaluate AI in clinical practice. First, clinical trials should include outcomes that reflect the success of implementation[89]. These include the compatibility with the workflow, the adoption rate, and the cost of implementation, which are crucial for understanding the success or failure of an AI tool.

Furthermore, with the current shortages in healthcare, an essential aspect of AI tools to consider during evaluation is the return-on-investment. A recent paper developed a return-on-investment calculator to inform decision making for an AI-powered radiology diagnostic imaging platform[90]. The calculator compares the current workflow to the updated workflow after deployment of the AI tool and provides insights into aspects such as time savings, effects on clinical outcomes, healthcare services provided, and the total cost. Of course, assumptions need to be made for aspects where data is unavailable. However, insights on the return-on-investment of AI tools will become increasingly important with a growing supply of tools.

Lastly, the introduction of LLMs poses considerable challenges to the evaluation of these tools due to their generative nature. Challenges include the lack of clearly defined evaluation metrics, variation in the design of human evaluations, and incorporating the human-LLM interaction into the evaluation[68]. In response to this, we recently published guidance, describing three tiers of clinical LLM validation (see Box 1)[91]. There are two recent examples of rigorous evaluations of LLMs in healthcare. The first study evaluated the value of

a tool that automatically generates notes from clinical conversations[23]. The tool was implemented in clinical practice and available for 10.000 physicians and staff. The authors evaluated the output of the tool automatically and manually, investigated the effect of the tool on several EHR logging metrics (e.g. time spent in the EHR after working hours), and collected patient and physician experiences. The second study evaluated several adapted LLMs on four clinical summarization tasks[92]. Apart from performing similar extensive automatic and manual evaluations of the summaries, they also investigated the potential medical threat that errors could pose. These two studies provide important examples on how to rigorously evaluate LLMs in clinical settings.

Box 1. Three tiers of medical Large Language Model validation[91]

> **1. General validation**
> General validation assesses general LLM quality independent of the performed task. Important outcomes at this stage may be the LLM's robustness to different formulations of the same prompt and the readability of the LLM output.
> **2. Task specific validation**
> Task specific validation assesses the LLM performance on task specific outcomes. For example, for summarization it may be the consistency with source material and coverage of important clinical concepts.
> **3. Clinical validation**
> Clinical validation assesses the LLM performance on specific healthcare outcomes. The validation goals at this tier will depend on the clinical objectives and intended use, such as improved health outcomes, higher patient satisfaction or reduction in administration time.

9.4 Future outlook

9.4.1 Large language models

The field of natural language processing has changed inconceivably over the past few years. Recent studies describe the promising performance of LLMs

in extracting information, drafting responses to patients, summarizing patient records, and even for prediction tasks[68,93]. A notable breakthrough was reported in a recent study by Google, where an AI system was capable of conducting a comprehensive diagnostic (chat) dialogue with patients, achieving diagnostic accuracy that surpassed that of trained physicians[83]. While LLMs have demonstrated significant potential, the translation of these capabilities into practical healthcare benefits remains largely unproven. Moreover, LLMs introduce several new challenges that need attention. These challenges include their generative nature and the hallucinations this may lead to, the energy consumption necessary for training and inference, and concerns about data privacy[64,94]. The latter might be even more pressing than before, because of the accessibility of tools such as ChatGPT. Educating healthcare professionals on the basics of AI and how they might use these tools is crucial to ensure appropriate and safe use.

Given these challenges and the current uncertainties inherent in new technologies, we should keep a critical stance towards the use of LLMs while also recognizing the potential opportunities they may offer. There are some interesting examples of tools that use LLMs as part of their NLP pipeline, using their language generation capabilities in combination with information retrieval models or curated datasets to control the possible outputs[95,96]. As stated in Section 9.3.4, such tools should be rigorously evaluated before use in clinical practice. It is clear, however, that the introduction of large language models has completely changed the field of NLP and will greatly impact how we practice healthcare in the coming years.

9.4.2 Role of companies

As discussed in our scoping review (Chapter 2), the current role of commercial companies within the field is ambivalent. On one hand, companies are often not transparent about the development and validation of their models. On the other hand, they play an essential role in turning promising technologies into software applications that can be used in clinical practice. With legislation such as the EU's Medical Device Regulation and the AI-act, this process is lengthy and expensive. Furthermore, during this process input from many different experts is needed.

To make sure promising NLP models end up as reliable software products, good collaboration between researchers and companies is needed. Creating a pipeline where healthcare organizations work on proof-of-concepts, which, if successful, are further developed into software products by companies could be beneficial to all parties involved. Validation and impact assessment by independent researchers should be a mandatory part of this pipeline, recognizing the dynamic nature of the field with continuous improvements.

Furthermore, electronic health records (EHR) vendors should play an important role in these developments. Since the launch of ChatGPT, EHR vendor Epic has greatly invested in integrating this technique into their EHR[97]. However, the largest Dutch EHR vendor, has yet to facilitate easy integration of AI. Governments and healthcare organizations should actively think about what this landscape should look like and the accompanying demands this should place on EHR vendors.

9.5 Recommendations

For hospitals	For developers
Develop a strategy for AI in general. This should include objectives for the use of AI, roles and responsibilities for development and deployment, and policy on how and when to engage with companies.	Be steered by the needs from healthcare, instead of the availability of data.
Invest in robust infrastructure to facilitate the deployment of NLP tools.	Include clinical expertise from the start to get in-depth knowledge about the meaning of the data and the clinical workflow.
Include all relevant expertise for every development and implementation project. If you do not have this expertise internally, make sure to involve independent external experts.	Critically reflect on the necessity to use complex NLP models. Sometimes less is more.
Educate your healthcare professionals on the basics of LLMs, how to use them responsibly, and the associated risks.	Rigorously evaluate your NLP models using established reporting standards and evaluation frameworks.

9.6 Conclusion

The field of natural language processing has seen inconceivable progress over the past five years, and with it the possibilities for application in healthcare. Although we identify many promising applications in this thesis, challenges related to data quality and availability, bias, and a lack of insights in clinically relevant metrics remain. These challenges hinder the further development or implementation of many NLP models in healthcare. To turn NLP models into valuable additions for clinical practice, we should pay more attention to working on the right problems, reporting on clinically relevant metrics, lowering the engineering effort needed to integrate a model into the clinical workflow, and performing thorough clinical impact evaluations. If these challenges are addressed, NLP may significantly improve clinician experience, patient experiences and outcomes, reduce costs, and keep healthcare accessible and affordable.

References

1. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, Turner K. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Comput Biol Med 2023;155:106649. PMID:36805219
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv 2019; PMID:31501885
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. Arxiv 2017;
4. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Arxiv 2018;
5. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv 2023; doi: 10.48550/arxiv.2307.09288
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv 2020; doi: 10.48550/arxiv.2005.14165
7. Anjum A, Zhao X, Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. Proc 3rd Ieee Acm Int Conf Big Data Comput Appl Technologies 2016;94–99. doi: 10.1145/3006299.3006335
8. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. Jmir Medical Informatics 2018;6(4):e45. PMID:30497991
9. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, Asch DA, Ungar LH, Merchant RM. Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. Health Affair 2017;35(4):697–705. PMID:27044971
10. Cammel SA, Vos MSD, Soest D van, Hettne KM, Boer F, Steyerberg EW, Boosman H. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. Bmc Med Inform Decis 2020;20(1):97. PMID:32460734
11. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. Bmj Heal Care Informatics 2021;28(1):e100262. PMID:33653690

12. Jiang H, Huang X, Zhang J, Song Z, Toral XS, Xu Y, Liu A, Guo L, Powell G, Verma A, Buckeridge D, Marelli A, Li Y. Supervised multi-specialist topic model with applications on large-scale electronic health record data. Proc 12th ACM Conf Bioinform, Comput Biol, Heal Inform 2021;1–26. doi: 10.1145/3459930.3469543
13. Ahuja Y, Zou Y, Verma A, Buckeridge D, Li Y. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. J Biomed Inform 2022;134:104190. PMID:36058522
14. Grootendorst M. LLM & Generative AI. Available from: https://maartengr.github.io/BERTopic/getting_started/representation/llm.html [accessed May 3, 2024]
15. Mu Y, Dong C, Bontcheva K, Song X. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. arXiv 2024; doi: 10.48550/arxiv.2403.16248
16. Williams CYK, Bains J, Tang T, Patel K, Lucas AN, Chen F, Miao BY, Butte AJ, Kornblith AE. Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries. medRxiv 2024;2024.04.03.24305088. PMID:38633805
17. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Heal 2023;5(4):e179–e181. PMID:36894409
18. Roberts RHR, Ali SR, Dobbs TD, Whitaker IS. Can Large Language Models Generate Outpatient Clinic Letters at First Consultation That Incorporate Complication Profiles From UK and USA Aesthetic Plastic Surgery Associations? Aesthetic Surg J Open Forum 2023;6:ojad109. PMID:38192329
19. Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, Jiang X, Guo L, Cai X, Zhang S, Zhang T, Zhu D, Shen D, Liu T, Li X. ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. arXiv 2023; doi: 10.48550/arxiv.2304.08448
20. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142. PMID:38413730
21. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study. Int J Méd Inform 2024;187:105443. PMID:38615509
22. Yim W, Fu Y, Abacha AB, Snider N, Lin T, Yetisgen M. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. Sci Data 2023;10(1):586. PMID:37673893
23. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, Liu V, Lee K. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal 2024;5(3). doi: 10.1056/cat.23.0404
24. Lyu M, Peng C, Li X, Balian P, Bian J, Wu Y. Automatic Summarization of Doctor-Patient Encounter Dialogues Using Large Language Model through Prompt Tuning. arXiv 2024; doi: 10.48550/arxiv.2403.13089
25. Sezgin E, Sirrianni J, Kranz K. Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden. medRxiv 2023;2023.12.06.23299573. PMID:38106162

26. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B, Xu H. Deep learning in clinical natural language processing: a methodical review. J Am Méd Inform Assoc 2020;27(3):457–470. PMID:31794016

27. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, Taylor RA, Krumholz HM, Radev D. Neural Natural Language Processing for unstructured data in electronic health records: A review. Comput Sci Rev 2022;46:100511. doi: 10.1016/j.cosrev.2022.100511

28. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Méd Inform Decis Mak 2020;20(1):280. PMID:33121479

29. Mugisha C, Paik I. Pneumonia Outcome Prediction Using Structured And Unstructured Data From EHR. 2020 IEEE Int Conf Bioinform Biomed (BIBM) 2020;00:2640–2646. doi: 10.1109/bibm49941.2020.9312987

30. Hollander D den, Dirkson AR, Verberne S, Kraaij W, Oortmerssen G van, Gelderblom H, Oosten A, Reyners AKL, Steeghs N, Graaf WTA van der, Desar IME, Husson O. Symptoms reported by gastrointestinal stromal tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data. Support Care Cancer 2022;30(6):5137–5146. PMID:35233640

31. Dirkson A, Verberne S, Kraaij W, Oortmerssen G van, Gelderblom H. Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Sci Rep 2022;12(1):10317. PMID:35725736

32. Dirkson A, Verberne S, Oortmerssen G van, Gelderblom H, Kraaij W. How do others cope? Extracting coping strategies for adverse drug events from social media. J Biomed Inform 2023;139:104228. PMID:36309197

33. Jeyaraman M, Ramasubramanian S, Kumar S, Jeyaraman N, Selvaraj P, Nallakumarasamy A, Bondili SK, Yadav S. Multifaceted Role of Social Media in Healthcare: Opportunities, Challenges, and the Need for Quality Control. Cureus 2023;15(5):e39111. PMID:37332420

34. Aakre CA. Applying Natural Language Processing Neural Network Architectures to Augment Appointment Request Review of Self-Referred Patients to an Academic Medical Center. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2022;2022:85–91. PMID:35854757

35. Davoudi A, Lee NS, Luong T, Delaney T, Asch EL, Chaiyachati KH, Mowery DL. Identifying Medication-related Intents from a Bidirectional Text Messaging Platform for Hypertension Management: An Unsupervised Learning Approach. medRxiv 2021;2021.12.23.21268061. doi: 10.1101/2021.12.23.21268061

36. Huang M, Fan J, Prigge J, Shah ND, Costello BA, Yao L. Characterizing Patient-Clinician Communication in Secure Medical Messages: Retrospective Study. J Méd Internet Res 2022;24(1):e17273. PMID:35014964

37. Steitz BD, Sulieman L, Warner JL, Fabbri D, Brown JT, Davis AL, Unertl KM. Classification and analysis of asynchronous communication content between care team members involved in breast cancer treatment. JAMIA Open 2021;4(3):ooab049. PMID:34396056

38. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, Chipidza FE, Leeman J, Aerts HJWL, Miller T, Savova GK, Gallifant J, Celi LA, Mak RH, Lustberg M, Afshar M, Bitterman DS. The effect of using a large language model to respond to patient messages. Lancet Digit Heal 2024; PMID:38664108

39. Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, Osborne C, Diaz J, Desai N, Webb S, Polston G, Helsten T, Gross E, Thackaberry J, Mandvi A, Lillie D, Li S, Gin G, Achar S, Hofflich H, Sharp C, Millen M, Longhurst CA. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. JAMA Netw Open 2024;7(4):e246565. PMID:38619840

40. Edmondson ME, Reimer AP. Challenges Frequently Encountered in the Secondary Use of Electronic Medical Record Data for Research. CIN: Comput, Inform, Nurs 2020;38(7):338–348. PMID:32149742

41. Yim W-W, Wheeler AJ, Curtin C, Wagner TH, Hernandez-Boussard T. Secondary use of electronic medical records for clinical research: challenges and opportunities. Converg Sci Phys Oncol 2018;4(1):014001. PMID:29732166

42. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337–1340. PMID:31427808

43. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: An audit of problem list completeness during the COVID-19 pandemic. Int J Méd Inform 2021;150:104452. PMID:33864979

44. Wang EC-H, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. J Am Méd Inform Assoc 2020;27(8):1190–1197. PMID:32620950

45. Searle T, Ibrahim Z, Teo J, Dobson R. Estimating redundancy in clinical text. J Biomed Inform 2021;124:103938. PMID:34695581

46. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. JAMA Netw Open 2022;5(9):e2233348. PMID:36156143

47. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(suppl_1):D267–D270. PMID:14681409

48. Organization WH. ICD-10 : international statistical classification of diseases and related health problems : tenth revision. 2004. Available from: https://iris.who.int/handle/10665/42980 ISBN:9241546549

49. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Heal Technol Inform 2006;121:279–90. PMID:17095826

50. Oliveira JM de, Costa CA da, Antunes RS. Data structuring of electronic health records: a systematic review. Heal Technol 2021;11(6):1219–1235. doi: 10.1007/s12553-021-00607-w

51. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2(6):305–311. doi: 10.1038/s42256-020-0186-1

52. Guan J, Li R, Yu S, Zhang X. A Method for Generating Synthetic Electronic Medical Record Text. IEEEACM Trans Comput Biol Bioinform 2019;18(1):173–182. PMID:31647443

53. Zhou N, Wu Q, Wu Z, Marino S, Dinov ID. DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. J Méd Syst 2022;46(12):96. PMID:36380246

54. Shafran I, Du N, Tran L, Perry A, Keyes L, Knichel M, Domin A, Huang L, Chen Y, Li G, Wang M, Shafey LE, Soltau H, Paul JS. The Medical Scribe: Corpus Development and Model Performance Analyses. Arxiv 2020.

55. Oommen C, Howlett-Prieto Q, Carrithers MD, Hier DB. Inter-Rater Agreement for the Annotation of Neurologic Concepts in Electronic Health Records. medRxiv 2022;2022.11.16.22282384. doi: 10.1101/2022.11.16.22282384

56. Bajpai S, Bajpai RC, Chaturvedi HK. Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods. Journal of the Indian Academy of Applied Psychology 2015;41(No.3 (Special Issue)):20–27.

57. Gordon ML, Lam MS, Park JS, Patel K, Hancock J, Hashimoto T, Bernstein MS. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. CHI Conf Hum Factors Comput Syst 2022;1–19. doi: 10.1145/3491102.3502004

58. Raghu M, Blumer K, Sayres R, Obermeyer Z, Kleinberg R, Mullainathan S, Kleinberg J. Direct Uncertainty Prediction for Medical Second Opinions. arXiv 2018; doi: 10.48550/arxiv.1807.01771

59. Balagopalan A, Baldini I, Celi LA, Gichoya J, McCoy LG, Naumann T, Shalit U, Schaar M van der, Wagstaff KL. Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact. PLOS Digit Heal 2024;3(4):e0000474. PMID:38620047

60. Timmons AC, Duong JB, Fiallo NS, Lee T, Vo HPQ, Ahle MW, Comer JS, Brewer LC, Frazier SL, Chaspari T. A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. Perspect Psychol Sci 2023;18(5):1062–1096. PMID:36490369

61. Raza S, Garg M, Reji DJ, Bashir SR, Ding C. Nbias: A natural language processing framework for BIAS identification in text. Expert Syst Appl 2024;237:121542. doi: 10.1016/j.eswa.2023.121542

62. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB, Foundation W DC Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging, Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. npj Digit Med 2023;6(1):170. PMID:37700029

63. Johnson RL, Pistilli G, Menédez-González N, Duran LDD, Panai E, Kalpokiene J, Bertulfo DJ. The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv 2022; doi: 10.48550/arxiv.2203.07785

64. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. npj Digit Med 2023;6(1):135. PMID:37516790

65. Verbeek P-P, Tijink D. Guidance ethics approach: An ethical dialogue about technology with perspective on actions. ECP | Platform voor de InformatieSamenleving; 2020. Available from: https://ris.utwente.nl/ws/portalfiles/portal/247401391/060_002_Boek_Guidance_ethics_approach_Digital_EN.pdf [accessed May 2, 2024]

66. Hond AAH de, Buchem MM van, Hernandez-Boussard T. Picture a data scientist: a call to action for increasing diversity, equity, and inclusion in the age of AI. J Am Méd Inform Assoc 2022;29(12):2178–2181. PMID:36048021

67. Investigators A of URP. The "All of Us" Research Program. N Engl J Med 2019;381(7):668–676. PMID:31412182

68. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA 2023;330(9):866–869. PMID:37548965

69. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3(1):18. PMID:31592444

70. Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Méd Inform Assoc 2019;26(12):1655–1659. PMID:31192367

71. Abbasian M, Khatibi E, Azimi I, Oniani D, Abad ZSH, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O, Li L-J, Jain R, Rahmani AM. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digit Med 2024;7(1):82. PMID:38553625

72. Sande D van de, Genderen ME van, Huiskens J, Gommers D, Bommel J van. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensiv Care Med 2021;47(7):750–760. PMID:34089064

73. Porter ME, Teisberg EO. Redefining Health Care: Creating Value-based Competition on Results. Harvard Business Review Press; 2006. Available from: https://books.google.nl/books?id=cse2LOAndNIC ISBN:9781422133361

74. Bodenheimer T, Sinsky C. From Triple to Quadruple Aim: Care of the Patient Requires Care of the Provider. Ann Fam Med 2014;12(6):573–576. PMID:25384822

75. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf 2015;24(10):608. PMID:26038586

76. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, Carlasare LE, Dyrbye LN. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681–1694. PMID:30803733

77. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W-J, Sinsky CA, Gilchrist VJ. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Medicine 2017;15(5):419–426. PMID:28893811

78. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, Westbrook J, Tutty M, Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016;165(11):753. PMID:27595430

79. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, Wang W, Luft HS. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. Health Affair 2017;36(4):655–662. PMID:28373331

80. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, Torchiana DF. The Impact of Administrative Burden on Academic Physicians. Acad Med 2017;92(2):237–243. PMID:28121687

81. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–1940. PMID:37460753

82. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, Deng A, Baston A, Ross J, Idowu E, Teo JT, Dobson RJ. Foresight -- Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs. arXiv 2022; doi: 10.48550/arxiv.2212.08072

83. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S, Singhal K, Cheng Y, Hou L, Webson A, Kulkarni K, Mahdavi SS, Semturs C, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Karthikesalingam A, Natarajan V. Towards Conversational Diagnostic AI. arXiv 2024; doi: 10.48550/arxiv.2401.05654

84. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care. JAMA 2019;321(23):2281–2282. PMID:31107500

85. Faneyte S. Stappenplan Healthy AI (HAI). Maasstad Ziekenhuis; Available from: https://www.nvki.nl/community/threads/stappenplan-healthy-ai.471/ [accessed May 12, 2024]

86. Handelingsruimte A. Tool Handelingsruimte Waardevolle AI voor Gezondheid en Zorg. Dutch Ministry for Health, Welfare and Sport; Available from: https://nlaic.com/wp-content/uploads/2022/06/04a.-Hulpmiddel-Handelingsruimte-Waardevolle-AI-voor-gezondheid-en-zorg.pdf [accessed May 12, 2024]

87. Sectra Amplifier Services: AI adoption made easy. Available from: https://medical.sectra.com/product/sectra-amplifier-services/ [accessed May 12, 2024]

88. Blackford Analysis. Available from: https://blackfordanalysis.com [accessed May 12, 2024]

89. Sande D van de, Chung EFF, Oosterhoff J, Bommel J van, Gommers D, Genderen ME van. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. npj Digit Med 2024;7(1):58. PMID:38448743

90. Bharadwaj P, Nicola L, Breau-Brunel M, Sensini F, Tanova-Yotova N, Atanasov P, Lobig F, Blankenburg M. Unlocking the Value: Quantifying the ROI of Hospital AI. J Am Coll Radiol 2024; PMID:38499053

91. Hond A de, Leeuwenberg T, Bartels R, Buchem M van, Kant I, Moons KG, Smeden M van. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Heal 2024;6(7):e441–e443. PMID:38906607

92. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142. PMID:38413730

93. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388(25):2399–2400. PMID:37342941

94. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Proc 57th Annu Meet Assoc Comput Linguistics 2019;3645–3650. doi: 10.18653/v1/p19-1355

95. Consensus. AI Search Engine for Research. Available from: https://consensus.app/home/about-us/ [accessed May 13, 2024]

96. Glass.Health. AI-Powered Clinical Decision Support. Available from: https://glass.health [accessed May 13, 2024]

97. Center MN. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service. 2023. Available from: https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/ [accessed May 2, 2024]

Chapter 10

Summary

This thesis investigates the application of natural language processing (NLP) in healthcare. Part 1 describes the development of several NLP models in different healthcare settings and discusses its challenges. Part 2 focuses on determining the added value of NLP in clinical practice, discussing two pilot studies.

Chapter 2 provides a comprehensive scoping review and research agenda for the implementation of digital scribes in clinical practice. Digital scribes, which leverage automatic speech recognition (ASR) and NLP techniques, aim to alleviate clinician burnout by automating clinical documentation. It highlights that current research primarily focuses on technical validity without adequately addressing clinical validity, usability, or utility. Recommendations for future research include improving ASR accuracy, ensuring comprehensive validation, and enhancing transparency in reporting to bridge the gap between research and practical implementation.

Chapter 3 evaluates the use of NLP to predict acute care utilization (ACU) in oncology patients starting chemotherapy, using clinical notes. It compares deep learning models to manually engineered language features and structured health data (SHD) models. Results indicate that while SHD models slightly outperform NLP models, both approaches are viable. The study underscores the potential of NLP in clinical applications and highlights risk biases across diverse patient groups, suggesting that future research should focus on enhancing model generalizability and addressing these biases.

Chapter 4 discusses the development and validation of the Artificial Intelligence Patient-Reported Experience Measures (AI-PREM), a tool designed to automate the analysis of open-ended patient experience data using NLP (see Figure1). The AI-PREM encompasses a newly developed open-ended questionnaire, an NLP pipeline for analyzing responses through sentiment analysis and topic modeling, and a visualization interface for easy interpretation by healthcare professionals. The validation process affirmed the tool's capability to identify relevant patient experience themes and sentiments accurately, suggesting its applicability in clinical settings to support quality improvement efforts.

Chapter 5 presents the development of a model to identify depression concerns in cancer patients using NLP on patient messages. The study utilized messages from a secure patient portal at a cancer center, applying logistic regression, support vector machines, and two variations of BERT models (original and Reddit-pretrained) for classification. The best performance was achieved by the BERT models, indicating their effectiveness in detecting depression concerns. The study also explored model fairness and explainability, revealing performance disparities across demographic groups. This research underscores the potential of advanced NLP techniques in enhancing early depression detection in clinical settings, though it also highlights the need for careful consideration of inherent model biases and their impact on different patient groups.

Chapter 6 examines public perceptions of statins on Reddit, using AI to analyze 10,233 posts and comments. NLP models and clustering algorithms categorized the content into 100 topics across six thematic groups. The study suggests that AI can provide valuable insights into public opinions, helping healthcare professionals address misconceptions and improve statin adherence.

Chapter 7 investigates the AI-PREM tool's value in clinical practice, specifically for a vestibular schwannoma care pathway (see Figure 1). By comparing open-ended and closed-ended patient experience questionnaires, the study found that the AI-PREM provided more detailed patient feedback, identifying specific improvement areas that closed-ended PREMs missed. The findings highlight AI-PREM's potential to enhance patient-centered care through actionable feedback.

Chapter 8 examines the impact of automating clinical documentation with large language models, focusing on time and quality (see Figure 2). Medical students summarized 150 mock conversations either manually or using a Dutch digital scribe system, finding that fully automated summaries had lower quality scores compared to manual ones, but editing of automated summaries improved in quality. Furthermore, manual summarization took longer than editing automated summaries. The study suggests digital scribes can reduce documentation time while maintaining high-quality records, with further research needed in clinical settings.

In conclusion, there is significant progress in NLP with many potential applications in healthcare. Despite these advancements, challenges such as data quality, bias, clinically relevant metrics, and the lack of high-quality clinical evaluations hinder widespread adoption. Focused efforts are required, such as more clinical evaluations, improved reporting, and more streamlined integration of NLP into clinical practice. This way, the potential of NLP to improve patient outcomes, enhance clinician experience, and reducing costs may be realized.

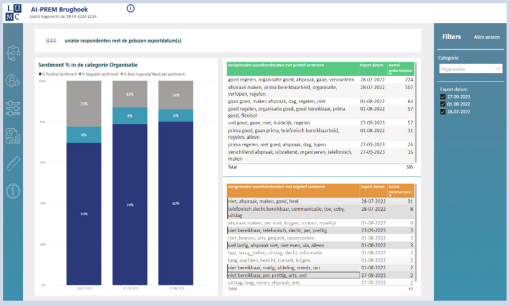
AI-PREM

Goal: Getting insights from free text patient experiences.

End user: Care pathway team, responsible for the quality improvement of the care pathway (varies per care pathway).

Workflow: Every year, the care pathway sends out a questionnaire consisting of five open-ended questions to all patients within the care pathway. The care pathway team then analyzes the AI-PREM results, specifically examining negative clusters of patient experiences to identify areas for improvement. Relevant topics are subsequently investigated by the team by examining the raw text data within those topics. When potential actions are identified, they are deliberated in meetings and balanced against any positive feedback on the same issues.

Application: Integrated as extra tab within the already existing Patient Reported Outcome Measures Dashboard.



Screenshot of the current version of the AI-PREM. On the left, the output from the sentiment analysis is shown per year. On the right, the topics per question are shown.

Figure 1: a description of the AI-PREM tool.

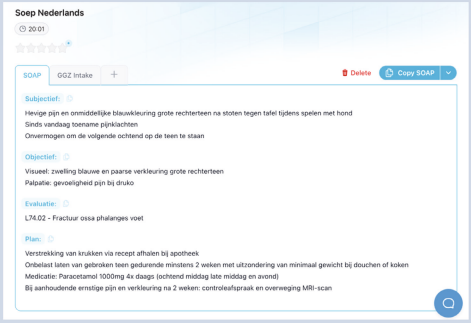
Autoscriber

Goal: Automatic note generation for clinical consultations.

End user: Physicians.

Setting: At the start of a consultation, the physician asks the patient if they consent with using Autoscriber. If they do, the physicians starts a new recording and proceeds as normal. In the back, Autoscriber automatically transcribes the conversation. At the end of the consultation, the physician ends the recording and after a few seconds, an automatically generated summary appears. This summary can be edited and then transferred to the patient's medical file.

Application: Currently used as stand-alone web application. However, integration with Hix (the largest EHR within the Netherlands) will soon be possible.



Screenshot of the current version of Autoscriber, showing an automatic summary of a mock conversation.

Figure 2: a description of the Autoscriber tool.

Chapter 11

Samenvatting

Dit proefschrift onderzoekt de toepassing van natuurlijke taalverwerking (NLP) in de gezondheidszorg. Deel 1 beschrijft de ontwikkeling van verschillende NLP-modellen in diverse zorgsettings en bespreekt de bijbehorende uitdagingen. Deel 2 richt zich op het vaststellen van de toegevoegde waarde van NLP in de klinische praktijk en bespreekt twee pilotstudies.

Hoofdstuk 2 biedt een uitgebreide scoping review en een onderzoeksagenda voor de implementatie van *digital scribes* in de klinische praktijk. Digital scribes, die gebruik maken van automatische spraakherkenning (ASR) en NLP-technieken, zijn bedoeld om de werkdruk van clinici te verlichten door de klinische documentatie te automatiseren. Onze review benadrukt dat huidig onderzoek voornamelijk gericht is op technische validiteit, zonder voldoende aandacht te besteden aan klinische validiteit, bruikbaarheid of nut. Aanbevelingen voor toekomstig onderzoek omvatten het verbeteren van de ASR-nauwkeurigheid, het waarborgen van uitgebreide validatie en het verbeteren van de transparantie in rapportage om de kloof tussen onderzoek en praktische implementatie te overbruggen.

Hoofdstuk 3 evalueert het gebruik van NLP om het gebruik van acute zorg bij oncologiepatiënten die starten met chemotherapie te voorspellen, door gebruik te maken van klinische notities. Het vergelijkt deep learning-modellen met modellen die gebruik maken van handmatig geëxtraheerde taalfeatures en gestructureerde gezondheidsgegevens (SHD). De resultaten geven aan dat, hoewel SHD-modellen iets beter presteren dan NLP-modellen, beide benaderingen rendabel zijn. De studie benadrukt het potentieel van NLP in klinische toepassingen en wijst op risico bias over diverse patiëntengroepen, wat suggereert dat toekomstig onderzoek zich moet richten op het verbeteren van de generaliseerbaarheid van modellen en het aanpakken van deze bias.

Hoofdstuk 4 bespreekt de ontwikkeling en validatie van de Artificial Intelligence Patient-Reported Experience Measures (AI-PREM), een tool die is ontworpen om de analyse van vrije tekst patiëntervaringsgegevens te automatiseren met behulp van NLP (zie Figuur 1). De AI-PREM omvat een nieuw ontwikkelde vragenlijst met open vragen, een NLP-pijplijn voor het analyseren van reacties via sentimentanalyse en topic modeling, en een visualisatie-interface voor gemak-

kelijke interpretatie door zorgprofessionals. Het validatieproces bevestigde de mogelijkheid van de tool om relevante patiëntervaringsthema's en -sentimenten nauwkeurig te identificeren, wat de toepasbaarheid ervan suggereert ter ondersteuning van kwaliteitsverbetering in klinische settings.

Hoofdstuk 5 presenteert de ontwikkeling van een model om mogelijke depressie bij kankerpatiënten te identificeren met behulp van NLP op patiëntberichten. De studie maakte gebruik van berichten van een beveiligd patiëntenportaal in een kankercentrum, waarbij gebruik werd gemaakt van logistische regressie, support vector machines en twee varianten van BERT-modellen (origineel en voorgetraind op Reddit) voor classificatie. De beste prestaties werden bereikt door de BERT-modellen. Dit geeft een behoorlijk goede effectiviteit aan in het detecteren van mogelijke depressie. De studie onderzocht ook de rechtvaardigheid en verklaarbaarheid van modellen, waarbij prestatieverschillen tussen demografische groepen aan het licht kwamen. Dit onderzoek onderstreept het potentieel van geavanceerde NLP-technieken voor vroege depressiedetectie in klinische settings, hoewel het ook de noodzaak benadrukt om bias en hun impact op verschillende patiëntengroepen zorgvuldig te overwegen.

Hoofdstuk 6 onderzoekt de publieke percepties van statines op Reddit, door met AI 10.233 posts en commentaren te analyseren. NLP-modellen en clusteringalgoritmen categoriseerden de inhoud in 100 onderwerpen over zes thematische groepen. De studie suggereert dat AI waardevolle inzichten kan bieden in publieke opinies, waardoor zorgprofessionals misvattingen kunnen aanpakken en de naleving van statinegebruik kunnen verbeteren.

Hoofdstuk 7 onderzoekt de waarde van de AI-PREM-tool in de klinische praktijk, specifiek voor het brughoektumorzorgpad (zie Figuur 1). Door open en gesloten patiëntervaringsvragenlijsten te vergelijken, vond de studie dat de AI-PREM meer gedetailleerde patiëntfeedback bood en specifieke verbeterpunten identificeerde die gesloten vragenlijsten misten. De bevindingen benadrukken het potentieel van AI-PREM om patiëntgerichte zorg te verbeteren door middel van actiegerichte feedback.

Hoofdstuk 8 onderzoekt de impact van het automatiseren van klinische documentatie met grote taalmodellen, met de nadruk op tijd en kwaliteit (zie Figuur 2). Geneeskundestudenten hebben 150 gesimuleerde gesprekken samengevat, hetzij handmatig, hetzij met een Nederlands digital scribe-systeem. Zij ontdekten dat volledig geautomatiseerde samenvattingen lagere kwaliteitsscores hadden vergeleken met handmatige samenvattingen, maar dat de kwaliteit verbeterde na het bewerken van geautomatiseerde samenvattingen. Bovendien kostte handmatig samenvatten meer tijd dan het bewerken van geautomatiseerde samenvattingen. De studie suggereert dat digital scribes de documentatietijd kunnen verminderen terwijl de kwaliteit van de verslagen hoog blijft, met verder onderzoek nodig in klinische omgevingen.

Concluderend: er is aanzienlijke vooruitgang in natural language processing (NLP) met vele potentiële toepassingen in de gezondheidszorg. Ondanks deze vooruitgang belemmeren uitdagingen zoals datakwaliteit, bias, klinisch relevante maten, en gebrek aan goede klinische studies de brede acceptatie. Er zijn gerichte inspanningen vereist, zoals meer klinische evaluaties, verbeterde rapportage en meer gestroomlijnde integratie van NLP in de klinische context. Dan kan het potentieel van NLP worden gerealiseerd voor het verbeteren van patiëntenuitkomsten, de ervaring van clinici en het verlagen van kosten.

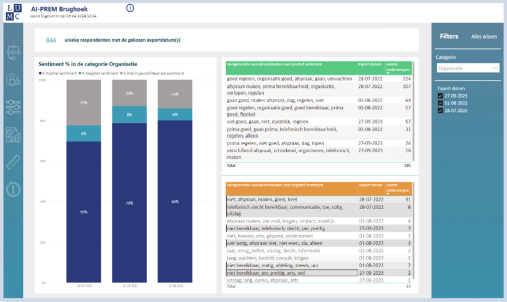
AI-PREM

Doel: Inzichten halen uit vrije tekst patiëntervaringen.

Eindgebruiker: Medewerkers van het zorgpad die verantwoordelijk zijn voor kwaliteitsverbeteringen.

Werkproces. Elk jaar stuurt het zorgpad een vragenlijst uit, bestaande uit vijf open vragen, naar alle patiënten die afgelopen jaar een afspraak hebben gehad. De reacties worden door de AI-PREM geanalyseerd en gevisualiseerd in een dashboard. Het zorgpadteam bekijkt het dashboard, met specifieke aandacht voor de negatieve onderwerpen om mogelijke verbeterpunten op te halen. Van relevante onderwerpen worden ook de onbewerkte reacties van patiënten bekeken. Potentiële verbeterpunten worden in meetings besproken en afgezet tegen positieve feedback op hetzelfde onderwerp.

Applicatie: Geïntegreerd als extra tab binnen het bestaande waardegedreven zorg dashboard.



Screenshot van de huidige versie van de AI-PREM. Links worden de uitkomsten van de sentiment analyse getoond per jaar, rechts staan de onderwerpen per vraag.

Figuur 1: een beschrijving van de AI-PREM tool.

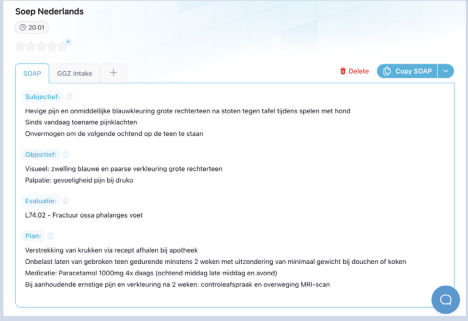
Autoscriber

Doel: Automatische generatie van de klinische verslaglegging voor consulten.

Eindgebruiker: Artsen.

Setting: Aan het begin van een consult vraagt de arts aan de patiënt of ze toestemming geven voor het gebruik van Autoscriber. Hierna start de arts een nieuwe opname en vervolgt het consult zoals gewoonlijk. Op de achtergrond transcribeert Autoscriber het gesprek. Aan het einde van het consult stopt de arts de opname en verschijnt na een paar seconde een automatisch gegenereerde samenvatting. Deze samenvatting kan aangepast worden en overgezet worden naar het dossier van de patiënt.

Applicatie: Op dit moment als losse webapplicatie. Binnenkort is integratie met Hix (het meest gebruikte elektronisch patiëntendossier in Nederland) mogelijk.



Screenshot van de huidige versie van Autoscriber, waarop een automatisch gegenereerde samenvatting van een testgesprek te zien is.

Figuur 2: een beschrijving van de Autoscriber tool.

Appendices

Publications

Curriculum vitae

Dankwoord

Publications

1. M.P. Bauer, **M.M. van Buchem** & S.A. Cammel. [Clinical history in times of big data; a plea for a standard for the structured recording of the clinical history]. *Ned. Tijdschr. voor Geneeskd.* 164**,** (2020).

2. **M.M. van Buchem**, H. Boosman, M.P. Bauer, I.M.J. Kant, S.A. Cammel & E.W. Steyerberg. The digital scribe in clinical practice: a scoping review and research agenda. *Npj Digital Medicine* 4, 57 (2021).

3. C. Fanconi, **M.M. van Buchem** & T. Hernandez-Boussard. Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes. *AMIA Jt Summits Transl Sci Proc*. eCollection 2023, 138-147 (2023).

4. A.A.H. de Hond, **M.M. van Buchem** & T. Hernandez-Boussard. Picture a data scientist: a call to action for increasing diversity, equity, and inclusion in the age of AI. *J Am Med Inform Assoc*. 29, 2178–2181 (2022).

5. **M.M. van Buchem**, O.M. Neve, I.M.J. Kant, E.W. Steyerberg, H. Boosman & E.F. Hensen. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM). *BMC Med Inform Decis Mak*. 22, 183 (2022).

6. **M.M. van Buchem**, H.H. 't Hart, P.J. Mosteiro, I.M.J. Kant & M.P. Bauer. Diagnosis Classification in the Emergency Room Using Natural Language Processing. In: *Caring is Sharing – Exploiting the Value in Data for Health and Innovation*. Vol. 302 815–816 (European Federation for Medical Informatics (EFMI) and IOS Press, 2023).

7. S. Somani, **M.M. van Buchem**, A. Sarraju, T. Hernandez-Boussard & F. Rodriguez. Topics and Sentiments around Statins on Reddit Using Artificial Intelligence. *Journal of the American College of Cardiology*. 81(8_Supplement), 1637-1637 (2023).

8. S. Somani**, M.M. van Buchem**, A. Sarraju, T. Hernandez-Boussard & F. Rodriguez. Artificial Intelligence–Enabled Analysis of Statin-Related Topics and Sentiments on Social Media. *JAMA Netw Open* 6**,** e239747 (2023).

9. O.M. Neve, **M.M. van Buchem**, M. Kunneman, P.P.G. van Benthem, H. Boosman & E.F. Hensen. The added value of the artificial intelligence patient-reported experience measure (AI-PREM tool) in clinical practise: Deployment in a vestibular schwannoma care pathway. *PEC Innov.* 3, 100204 (2023).

10. M.P. Bauer, **M.M. van Buchem** & S. Verberne. [Making a diagnosis with the aid of the internet in times of large language models]. *Nederlands Tijdschrift Voor Geneeskunde.* 167: D7998-D7998 (2023).

11. A.A.H. de Hond, **M.M. van Buchem**, C. Fanconi, M. Roy, D. Blayney, I.M.J. Kant, E.W. Steyerberg & T. Hernandez-Boussard. Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study. *JMIR Med Inform.* 12, e51925 (2024).

12. A.A.H. de Hond, T. Leeuwenberg, R. Bartels**, M.M. van Buchem**, I.M.J. Kant, K.G.H. Moons & M. van Smeeden. From text to treatment: The crucial role of validation for generative Large Language Models in healthcare. *Lanc Digit Health*. 6, e441-e443 (2024).

13. M.M. van Buchem, A.A.H. de Hond, C. Fanconi, V. Shah, M. Schuessler, I.M.J. Kant, E.W. Steyerberg & T. Hernandez-Boussard. Applying Natural Language Processing to Patient Messages to Identify Depression Concerns in Cancer Patients. *J Am Med Inform Assoc*. ocae188 (2024).

14. M.M. Rakers MM, **M.M. van Buchem**, S. Kucenko, A.A.H. de Hond, I.M.J. Kant, M. van Smeeden, K.G.M. Moons, A.M. Leeuwenberg, N. Chavannes, M. Villalobos-Quesada, H.J.A. van Os. Availability of Evidence for Predictive Machine Learning Algorithms in Primary Care: A Systematic Review. *JAMA Netw Open*. 7(9):e2432990 (2024).

15. **M.M. van Buchem**, I.M.J. Kant, L. King, J. Kazmaier, E.W. Steyerberg & M.P. Bauer. Automating Clinical Documentation Using Large Language Models: Effects on Documentation Time and Quality. *JMIR AI.* 3;e60020 (2024).

Curriculum vitae

Marieke Meija van Buchem werd in 1995 geboren te Princeton, Verenigde Staten. In 2013 behaalde zij haar VWO-diploma aan het Stedelijk Gymnasium te Leiden. Ze ging Geneeskunde studeren aan de Vrije Universiteit van Amsterdam en behaalde in 2016 haar bachelordiploma. Als onderdeel van de master Geneeskunde liep ze een wetenschappelijke stage bij het Alzheimercentrum. Daar werd haar interesse in technologie en innovatie in de zorg versterkt, en ze besloot een overstap te maken naar de 2-jarige master Medical Informatics aan de Universiteit van Amsterdam. Voor haar masterscriptie kon ze terecht in het LUMC, bij het pas opgerichte AI-team, waar ze werkte aan de allereerste versie van een digital scribe (nu de startup Autoscriber). Na het cum laude behalen van haar master in 2020 begon ze direct aan haar PhD in het LUMC. Hier richtte ze zich op het toepassen van natuurlijke taalverwerking in het ziekenhuis. In 2022 werd ze toegelaten tot het internationale Google Cloud Research Innovators-programma, en bracht ze daarnaast zes maanden door aan Stanford University. Na haar terugkeer in Nederland werkte Marieke, naast haar PhD, aan het verder professionaliseren en uitbreiden van het AI-team en het CAIRELab, het AI-expertisecentrum van het LUMC. In 2023 kreeg ze voor haar brede bijdrage aan het veld de Young Professional Award van Women in AI. Sinds 2023 is Marieke officieel werkzaam als innovatiemanager in het AI-team. Hier werkt ze aan de strategie en organisatie om de waarde van AI in de zorg te realiseren.

Dankwoord

Mijn promotie is allesbehalve voorspelbaar geweest, zowel op persoonlijk als op professioneel en vakinhoudelijk gebied. Na één maand brak corona uit, na anderhalf jaar werd mijn team grotendeels wegbezuinigd en na tweeënhalf jaar zorgde de introductie van ChatGPT voor een complete revolutie in mijn onderzoeksveld en ver daarbuiten. Dit alles heeft geleid tot een ontzettend dynamisch, interessant promotietraject, waarin ik me zowel op persoonlijk als op professioneel vlak enorm heb kunnen ontwikkelen. Er zijn ontzettend veel mensen om hiervoor te bedanken.

Allereerst mijn (co)promotoren: Ewout, ik ging altijd weer met meer kennis bij jou weg. Dank voor alle goede discussies en gesprekken, of ze nou over AI, CAIRELab of Amerika gingen. Ilse, jij hield me goed op koers en bracht richting als ik het overzicht verloor. Dat is heel belangrijk voor me geweest. Martijn, van stagebegeleider tot copromotor: met jou werk ik het allerlangst samen. Ik waardeer onze samenwerking enorm.

Dan mijn eerste AI-team: Maurice, Simone, Marjolein, Esmee, Charlotte, Feline, Laurens, Ilse en Anne. Ik kijk met enorm veel plezier terug op onze vele lunchrondjes, teamborrels (online en offline) en heidagen. Ik realiseer me steeds meer hoe bijzonder onze pioniersrol is geweest en met wat voor positiviteit en energie we die uitdaging zijn aangegaan.

Dan veel mensen waarmee ik heel fijn heb samengewerkt: Erik, Olaf en Hileen, hét voorbeeld van succesvolle (en gezellige) cocreatie. Simone en Martijn, aan ambitie geen gebrek. Vóór de introductie van LLM's al proberen om een digital scribe te bouwen bleek een enorme uitdaging. Ik heb altijd veel energie uit dit project gehaald; en nog steeds, nu samen met de collega's van Autoscriber. Tina, your energy and collaborative mindset have been an inspiration to me. Thank you for hosting and mentoring me during my time at Stanford. Claudio, I still miss our tea time sessions, where we talked about everything from personal pet peeves to technical deep dives. You really gave color to my time at Boussard Lab. Suzan, ik heb ontzettend veel gehad aan ons contact. Dank voor de gastvrijheid

waarmee je me uitnodigde voor jouw labmeetings, interessante congressen en gezamenlijke artikelen en praatjes.

Floor, Siri, alhoewel jullie op andere plekken in het LUMC promoveerden, waren de struggles hetzelfde. Dan moet je jezelf vooral niet te serieus nemen en lekker gaan fietsen, rennen of schaatsen om alles weer op een rijtje te krijgen.

Lieve Anne, mijn PhD-buddy vanaf dag één: zonder jou was het allemaal niks geworden! Er waren genoeg uitdagingen tijdens onze PhD, stuk voor stuk zijn we deze samen aangegaan. En dan onze grote reis: samen naar de Bay Area verhuizen, waar we als twee veel te vrolijke enthousiastelingen Boussard Lab kwamen opfleuren. Dank voor je onvoorwaardelijke steun en vriendschap, we made it.

De laatste periode van mijn promotie was ik vooral druk bezig met het opzetten van een nieuw AI-team. Jeroen, jij hebt mij een nieuwe functie aangeboden terwijl ik mijn promotie nog moest afronden. Jouw impliciete vertrouwen in mijn kunnen heeft me enorm gesterkt, zowel tijdens mijn promotie als daarna. Alexander, Marijke, Joris, Leandra en Masoumeh: ontzettend bedankt voor jullie steun en aanmoediging om, ondanks mijn constante gebrek aan tijd, mijn promotie af te ronden.

Dan de mensen die zorgden voor een gezonde work-life balance: al mijn lieve vrienden en vriendinnen uit Amsterdam, Leiden, van HealthInnovaitors, en iedereen daaromheen.

Mijn fantastische ooms, tantes, neven, nichten en lieve schoonfamilie: ik haal veel inspiratie uit onze hechte, diverse familie. Dank voor alle gezelligheid en steun.

Opappa en Omamma, jullie hebben de trend gezet met academische verblijven in Amerika. Bijna 40 jaar nadat jullie naar Stanford verhuisden, trad ik in jullie voetsporen. Heel bijzonder om al deze ervaringen met jullie te kunnen delen. Opa, ik heb vaak gedacht aan jouw uitspraak: gewoon de ene voet voor de andere blijven zetten. Grootpappa en Grootmamma, jullie maken de afronding van mijn

promotie niet meer mee, maar jullie onuitputtelijke energie, vertrouwen, interesse, eigenzinnigheid en autonomie waren en blijven als een lichtbaken.

Stijn en Brechtje, mijn squad. Ook zonder echt te weten wat ik de afgelopen jaren heb gedaan zijn jullie mijn grote steun en toeverlaat.

Pappa en Mamma, ik ben perongeluk precies uitgekomen op het snijvlak van jullie vakgebieden. Hoe kan het ook anders met zulke fantastische ouders. Jullie zien altijd meteen hoe het gaat en snappen wat ik nodig heb. Dank voor al jullie onvoorwaardelijke steun en vertrouwen. Jullie zijn mijn grote voorbeelden.

Lieve Tom, je zou bijna als medeauteur op de kaft mogen staan. Met je oneindige geduld, kopjes koffie, reality checks, avondwandelingetjes, mentale support, sportiviteit, structuur, vertrouwen en heel veel lol heb je me bijna letterlijk de streep over gekregen. Twee momenten springen eruit: je steun om zónder jou naar Stanford te vertrekken en je interventie een paar maanden geleden: Mariek, je moet je PhD gewoon áfmaken. Daar had je helemaal gelijk in, maatje.