Expert Systems With Applications

Automatic documentation of professional health interactions: a systematic review

--Manuscript Draft--

Manuscript Number:	ESWA-D-22-02520
Article Type:	Review article
Keywords:	EHR; Automatic documentation; Neural networks.
Corresponding Author:	Cristiano Andre André da Costa, Ph.D. Universidade do Vale do Rio dos Sinos São Leopoldo, BRAZIL
First Author:	Frederico Falcetta
Order of Authors:	Frederico Falcetta
	Fernando de Almeida
	Janaína Lemos
	José Roberto Goldim
	Cristiano André da Costa
Abstract:	Electronic systems are increasingly present in the healthcare system and are often related to improved medical care. However, the widespread use of these technologies ended up building a relationship of dependence that can disrupt the doctor-patient relationship itself. In this context, digital scribes are automated clinical documentation systems able to capture the physician–patient conversation and then generate the documentation for the appointment, enabling the physician to fully engage with the patient. We have performed a systematic literature review on intelligent solutions for automatic speech recognition (ASR) with automatic documentation during a medical interview. The scope included only original research on systems that could detect speech and transcribe it in natural and structured fashion simultaneously with the doctor-patient interaction, excluding speech-to-text only technologies. The search resulted in a total of 1995 titles, with 8 articles remaining after filtering for the inclusion and exclusion criteria. The intelligent models consisted mostly in an ASR system with natural language processing capability, a medical lexicon and a structured text output. None of the articles had a commercially available product at the time of the publication and very limited real-life experience was reported. So far none of the applications has been prospectively validated and tested in large scale clinical studies. Nonetheless, these first reports suggest that automatic speech recognition may be a valuable tool in the future to facilitate medical register in a faster and more reliable manner. By improving transparency, accuracy and empathy, it could drastically change the way patients and doctors experience a medical visit. Unfortunately, clinical data on usability and benefits of such applications is almost non-existent. We believe that future work in this area is necessary and needed.

São Leopoldo, Brazil, April 12th, 2022.

Dear Editor-in-chief,

We want to submit a possible contribution to the Expert Systems with Applications (ESWA) for your consideration. The article “Automatic documentation of professional health interactions: a systematic review” is original, and we are sending the manuscript to your journal only. The article explores recent literature related to intelligent solutions for automatic speech recognition (ASR) with automatic documentation during a medical interview. The article has identified the most relevant studies from the last ten years leading research indexers. Based on the analysis of the studies, we answer some research questions defined based on the authors’ knowledge of the subject. With the research questions, we could determine a taxonomy, recognize how the works are evaluated, identify challenges, and open questions to direct future work.

It is worth pointing out that this work reflects an active collaboration among professors in computer science and medicine. The researchers are from Universidade do Vale do Rio dos Sinos (São Leopoldo, Brazil), Hospital Fêmina (Porto Alegre, Brazil), and Hospital de Clínicas de Porto Alegre (Porto Alegre, Brazil). We think that this topic is of broad interest to the ESWA audience. If we can be of further assistance or information, please do not hesitate to contact us.

Best Regards,

Prof. Cristiano André da Costa, Ph.D. (Corresponding Author)

Address: SOFTWARELAB – Software Innovation Laboratory
Programa de Pós-graduação em Computação Aplicada (PPGCA)
Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos, 950
São Leopoldo, RS, Brazil, 93022-750
Phone: +55(51) 3590-8161 / +55(51) 35908162
E-mail: cac@unisinos.br / caccac@gmail.com

Automatic documentation of professional health interactions: a systematic review

Frederico Soares Falcetta, MD, M.Sc. in Pathology
Software Innovation Laboratory - SOFTWARELAB
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: fredfalcetta@gmail.com

Fernando Kude de Almeida, MD, M.Sc. in Pathology
Hospital Fêmina - Porto Alegre, Brazil
E-mail: fernandokude@gmail.com

Janaína Conceição Sutil Lemos, M.Sc. in Computer Science
Escola Politécnica
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: csutil@unisinos.br

José Roberto Goldim, Ph.D. in Medicine, Professor
Bioethics Division, Hospital de Clínicas de Porto Alegre/Brazil, Porto Alegre, Brazil.
E-mail: jgoldim@hcpa.edu.br

*Cristiano André da Costa, Ph.D. in Computer Science, Professor
Software Innovation Laboratory - SOFTWARELAB
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: cac@unisinos.br

***Corresponding author**

Software Innovation Laboratory - SOFTWARELAB

Programa de Pós-Graduação em Computação Aplicada

Universidade do Vale do Rio dos Sinos

Av. Unisinos 950 93022-000 São Leopoldo RS, Brazil
Phone: +55 51 35908161 Fax: +55 51 35908162

Acknowledgements

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel - CAPES (Finance Code 001) and the National Council for Scientific and Technological Development - CNPq (Grant Numbers 309537/2020-7 and 404572/2021-9) for supporting this work.

Automatic documentation of professional health interactions: a systematic review

Frederico Soares Falcetta, Fernando Kude de Almeida, Janaína Conceição Sutil Lemos, José Roberto Goldim and Cristiano André da Costa*

ARTICLE INFO

Keywords:
EHR
Automatic documentation
Neural networks

ABSTRACT

Electronic systems are increasingly present in the healthcare system and are often related to improved medical care. However, the widespread use of these technologies ended up building a relationship of dependence that can disrupt the doctor-patient relationship itself. In this context, digital scribes are automated clinical documentation systems able to capture the physician–patient conversation and then generate the documentation for the appointment, enabling the physician to fully engage with the patient. We have performed a systematic literature review on intelligent solutions for automatic speech recognition (ASR) with automatic documentation during a medical interview. The scope included only original research on systems that could detect speech and transcribe it in natural and structured fashion simultaneously with the doctor-patient interaction, excluding speech-to-text only technologies. The search resulted in a total of 1995 titles, with 8 articles remaining after filtering for the inclusion and exclusion criteria. The intelligent models consisted mostly in an ASR system with natural language processing capability, a medical lexicon and a structured text output. None of the articles had a commercially available product at the time of the publication and very limited real-life experience was reported. So far none of the applications has been prospectively validated and tested in large scale clinical studies. Nonetheless, these first reports suggest that automatic speech recognition may be a valuable tool in the future to facilitate medical register in a faster and more reliable manner. By improving transparency, accuracy and empathy, it could drastically change the way patients and doctors experience a medical visit. Unfortunately, clinical data on usability and benefits of such applications is almost non-existent. We believe that future work in this area is necessary and needed.

1. Introduction

Electronic systems are increasingly present in the healthcare system and often related to improved medical care. Doctors and nurses rely on Electronic Health Records (EHRs) and Electronic Medical Records (EMRs) to improve quality of patient care, analyze costs among other tasks. These systems contain data from the patient's medical record such as test results, information from their previous visits to other health professionals and their opinions (Heart, Ben-Assuli and Shabtai, 2017).

However, the widespread use of these technologies ended up building a relationship of dependence that can disrupt the doctor-patient relationship itself Pearce, Trumble, Arnold, Dwan and Phillips (2008). Before, the doctor's attention had few distractions and had its main focus on the patient in front of him. Currently, due to the ubiquity of computers and applications for recording the medical visit, the health professional's focus can be shifted or completely changed from the patient in front of him to electronic devices. In this context, a study showed that physicians spent from 24% to 55% of the medical interview looking at the computer screen, reducing their interaction with the patient and the possibility of showing emotional reaction to the patient's complaints (Margalit, Roter, Dunevant, Larson and Reis, 2006). The presence of a human scribe has been previously regarded as a possible solution for this problem and at least 20,000 of them were working in the United States by 2016, but concerns about costs and privacy with a third person in the room remain (Topol, 2019).

Some authors believe that the relationship has changed from dyad (doctor-patient) to a triad (doctor-computer-patient) (Scott and Purves, 1996). Furthermore, there is an erroneous impression that increasing the presence of technology would lead to a reduction in medical errors. Often, the standardization of treatment algorithms, copying of model reports and even medical records can lead to serious care errors. Electronic notes are often poorly readable, produced with overuse of copy and paste, and too large as a result of unfiltered data stored in other parts of the EHR.

ORCID(s):

In the context of inpatient care, physicians reported that writing progress notes in EHRs takes more time than it should and consequently the notes may not be available to other members of the care team at the end of the working day, causing negative impacts to patient care. In addition, the perception that electronic notes may not be accurate impairs also their use for research (Payne, Alonso, Markiel, Lybarger, Lordon, Yetisgen, Zech and White, 2018).

Attention and dependence on electronic devices is currently a popular topic and is associated with health problems such as anxiety and depression. A systematic review with meta-analysis identified that approximately a quarter of the population of children and young adults has problematic smartphone use with a behavior similar to addiction (Sohn, Rees, Wildridge, Kalk and Carter, 2019). Furthermore, the same study showed a strong correlation of this behavior with symptoms of depression, anxiety, and sleep problems. Another cross-sectional study assessed the prevalence of obsessive behavior in relation to the use of smartphones in a population of 688 university students in Lebanon and showed similar results such as abstinence, decreased sleep quality, and symptoms of depression and anxiety (Matar Boumosleh and Jaalouk, 2017).

For an electronic documentation system to be truly effective, it must eliminate or drastically reduce the need for the physician to split its attention by turning to a computer and manually perform some step to generate the documentation for the appointment. In this context, digital scribes are automated clinical documentation systems able to capture the physician–patient conversation and then generate the documentation for the appointment, enabling the physician to fully engage with the patient. By reducing the time and effort invested by physicians in the documentation process, digital scribes have the potential to improve the physician patient relationship and decrease physician burnout (Quiroz, Laranjo, Kocaballi, Berkovsky, Rezazadegan and Coiera, 2019). Furthermore, an automatically transcribed conversation could be more accurate and thorough than a regular manual transcription and allow patients to review and edit their own data (Topol, 2019).

In this context, the main contributions of this work are:

1. Present the intelligent models used in the the automatic documentation of health interactions
2. Describe the challenges to involved in the creation of an application for automatic documentation
3. Elucidate the problems involved with traditional documentation of the health interactions and how the automatic documentation could improve this process

This study is structured in 7 sections. In the second section we describe what constitutes the health record, the process called anamnesis or medical history taking and how this is one of the pillars of information gathering for diagnosis and treatment and the possibility of digital scribes to register this process automatically. The third section is a brief review of the field of artificial intelligence and applications that uses technologies such as artificial neural networks, deep learning and natural language processing. The fourth section deals with the methodology of this research paper and in the fifth section we describe the results of our survey. In the sixth section we discuss our results as well as future directions in this field of research. Finally, in the seventh section we present our conclusions.

Sections: Health records, automatic documentation, materials and methods

The second section is a brief review of the context Section 2 discusses the methods adopted in this survey and we present our results in Section 3. In Section 4, we answer and discuss our research questions and in Section 5, we explain some limitations of this work. Finally, Section 6 presents the final considerations of this review and its conclusions.

2. Health Records

Electronic Medical Records (EMRs) and Electronic Health Records (EHR) store medical records in a computerized way. In general, EMR is defined as an internal organizational system, while the EHR is usually defined as an inter-organizational system. Personal health records (PHR), in turn, are online systems used by patients. This type of system aims to give patients clear information and enable them to be engaged. If combined, these systems can enable all parties involved to be better informed, as well as their integration could help create massive databases that could be analyzed for research purposes (Heart et al., 2017).

EHR/EMR systems usually store patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports (Heart et al., 2017). It is important to note that EHR/EMR does not refer only to digitization of patient's medical data and eliminating the need for paper records. These systems aim to deliver quality care to patients and maintain their safety, as well as the integrity of their records (Aldosari, 2017).

However, EHR/EMR failed to achieve their goals entirely. A recent survey has found that 21.1% of the patients could find a mistake in their electronic health records and that they could be considered serious in 42.3% of the cases (Bell, Delbanco, Elmore, Fitzgerald, Fossa, Harcourt, Leveille, Payne, Stametz, Walker and DesRoches, 2020). Unfortunately, digitized systems permit that those errors be replicated indefinitely and degraded clinical documentation has also contributed to widespread physician dissatisfaction with EHRs (Payne et al., 2018; Hammond, Helbig, Benson and Brathwaite-Sketoe, 2003). At the same time, the adoption of EHRs is associated, at least in part, with increased physician burnout and stress Kroth, Morioka-Douglas, Veres, Babbott, Poplau, Qeadan, Parshall, Corrigan and Linzer (2019) and symptoms tend to get worse with increased after hours use and messaging (Adler-Milstein, Zhao, Willard-Grace, Knox and Grumbach, 2020).

As examples of determinants of EHR usage in developed countries can be mentioned design, ease of use, interoperability between departments, privacy and security, costs, user time/workload, organization size, IT support, and training. In this context, real-time testing of EMRs/EHRs in working environment combined with proper training from the vendors can contribute to reduce physician burnout due to the pressures of working with complex systems (Aldosari, 2017).

The process to gather information from the patient is called anamnesis or medical history taking. This conversation contains important information collected during a structured interview. It serves to assess the patient's complaint(s) with all its relevant characteristics and dimensions for a diagnosis to be made (Orient and Sapira, 2012). In addition to information related to the patient's complaints, past morbid history, family history and psychosocial characteristics are also collected and can serve to help the health professional to elaborate a set of diagnostic hypotheses (Stevenson, 1971). These diagnostic hypotheses constitute a list of diseases that may or may not be evaluated during the physical examination or through complementary exams, like laboratory or imaging tests. The final objective is to confirm one of the hypotheses and exclude the others, ending the diagnosis process.

Taking the clinical history is complex and requires training (Schnabel, 1983). This process can be divided into a practical and theoretical domain. In the domain of practical knowledge, we can mention the ability to collect information through the interview using specific techniques, such as silence, use of open questions, non-interruption of the patient while providing information, among others. The theoretical domain, on the other hand, constitutes the professional's entire clinical knowledge and can be summarized as the depth of his clinical knowledge (Weinstein, Fineberg, Elstein, Frazier, Neuhauser, Neutra and McNeil, 1980). The construction of a correct differential diagnosis therefore requires knowledge of each of the diseases that will be included and the types of information that the anamnesis will provide (Groopman, 2008).

During the anamnesis process, the multiple questions asked by the health professional build a dynamic list that will have items added and excluded depending on each of the patient's answers (Eisenberg, 1995). Anamnesis, therefore, is a dynamic process that can take different paths and depends solely on the questions asked and their answers (Sackett, 1992). In addition, the doctor-patient relationship plays a major role in facilitating this process and depends on other factors such as empathy exercised by the health professional, the patient's clinical situation, the environment where it's being carried out, among others (Lipkin, 1987). In this context, the health professional often records the anamnesis on the computer at the same time as he conducts it, saving time, but creating distractions leading to the loss of important information. Some professionals record the consult after the end of the consult, costing time and making the record prone to errors from the amount of detailed information that must be registered.

To enable an electronic record of the anamnesis without distracting or disturbing the doctor-patient relationship, new solutions can be devised. The use of automatic speech recognition and recording, which are based on the use of artificial intelligence is one of them. Below is a brief description of these technologies and their possibilities.

The information processing flow to generate documentation of patient-doctor interactions is shown in Figure 1.

As previously mentioned, digital scribes are automated systems able to capture the physician–patient conversation and then generate the documentation for the appointment. To generate medical notes for the clinician–patient encounter, a digital scribe must be able to: (1) record the clinician–patient conversation, (2) convert the audio to text, and (3) extract and summarize important information from the text (Quiroz et al., 2019).

The structure for a digital scribe includes a microphone that records a conversation, an Automatic Speech Recognition (ASR) system, that transcribes this conversation and Natural Processing Language (NLP) models. NLP is a branch of Artificial Intelligence (AI) that enables machines to understand the human language. NLP models are used to extract and/or summarize relevant information and present it to the physician. The extracted information can be used to create clinical notes or to support the diagnosis process. Several companies and startups are creating digital scribe systems (Quiroz et al., 2019; van Buchem, Boosman, Bauer, Kant, Cammel and Steyerberg, 2021; Ghatnekar, Faletsky

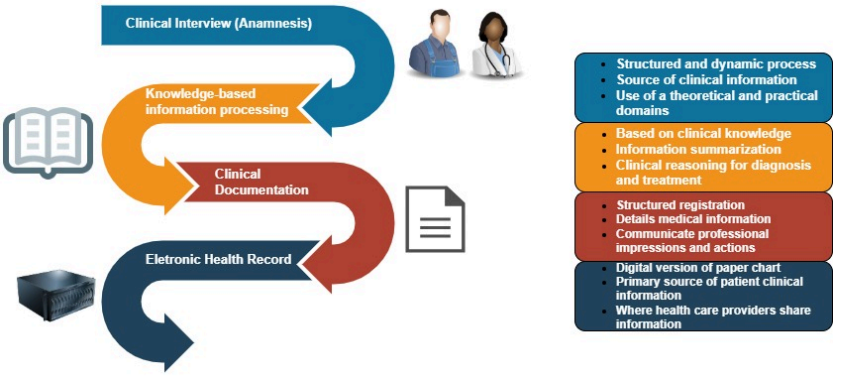

Figure 1: Information processing.

and Nambudiri, 2021). However, there are several concerns about implementing a digital scribe in healthcare. For example, the algorithms underlying digital scribe technologies must be trained to understand medical terminology and to adapt to different languages (and/or styles and accents in a specific language). In addition to the technical barriers to digital scribe implementation, there are both privacy and legal concerns (van Buchem et al., 2021; Ghatnekar et al., 2021).

A basic structure for digital scribes is shown in Figure 2.

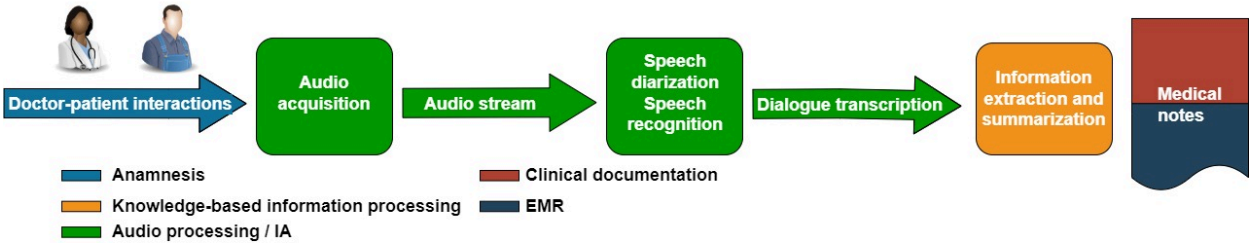

Figure 2: Structure for digital scribes

Because of the reasons presented above, successful implementation of a digital scribe requires a thorough investigation of its technical suitability as well as clinical validity, utility and usability (van Buchem et al., 2021).

3. Automatic Documentation

Artificial Intelligence (AI) aims to study, develop and employ machines to perform human activities autonomously. An AI solution involves several technologies that can simulate human capabilities related to intelligence, such as reasoning, perception of the environment and the ability to analyze for decision making. The large volumes of data available to organizations today enable the use of various AI resources, which was not possible decades ago (Paranjape, Schinkel, Panday, Car and Nanayakkara, 2019).

According to Panch, Szolovits and Atun (2018), the provision of health care involves basic tasks where significant amounts of data are processed: screening and diagnosis, where cases are classified based on history, examination and investigations, and subsequently, treatment and monitoring, which can be composed of several steps. The essential form of these processes in the domains of health system management and care delivery involves hypothesis generation, hypothesis testing and action. In this context, machine learning has the potential to improve hypothesis generation and testing within a healthcare system, revealing previously hidden trends in the data. AI has the advantage in relation to humans of being able to learn from millions of previous examples, what could increase diagnostic accuracy when

analysing images or lesions, help identify patients at risk of a particular disease or of clinical deterioration making clinical practice recommendations (Rajkomar, Dean and Kohane, 2019).

In addition to having the potential to assist, for example, in data analysis, disease diagnosis and treatment recommendation, AI can be applied to electronic health records, where specific algorithms are used to identify individuals with a family history of hereditary disease or increased risk of chronic diseases (Hamet and Tremblay, 2017; Jiang, Jiang, Zhi, Dong, Li, Ma, Wang, Dong, Shen and Wang, 2017). Further applications of AI in medicine go even beyond direct patients care and would encompass also medical research (big data analysis, drug discovery, multiple testing), home-care and self-diagnosis (Topol, 2019).

Some technologies that emerged in the field of AI and that have potential for application in medicine are deep learning and natural language processing. Deep learning is a specific type of machine learning that involves artificial neural networks, which are systems designed to function by sorting information in the same way as a human brain. In deep learning, neural networks with multiple layers of abstraction are used for pattern recognition and classification applications supported by datasets. The learning process takes place between its layers of mathematical neurons, in which information is transmitted by each layer. In this scheme, the output of the previous layer is the input of the posterior layer. Examples of applications include image identification and voice recognition (Abiodun, Jantan, Omolara, Dada, Mohamed and Arshad, 2018; Otter, Medina and Kalita, 2020).

Natural language processing, on the other hand, aims to study and try to reproduce the developmental processes linked to the functioning of human language. With this technology, machines can better understand texts – which involves recognizing context and extracting information. It is also possible to compose texts from data obtained by the computer (Otter et al., 2020).

As with any other technology, applying machine learning to medicine has met some challenges and limitations: access to good quality data (stored in different systems and in different formats); dependence on limited or biased information extracted from the patient by the staff physician; potential influence by incorrect prescribing habits and previous mistakes; over reliance on automatic diagnosis and procedures; and lack of real world prospective data on safety and outcomes with these new technologies (Rajkomar et al., 2019).

The speech recognition process transforms acoustic signals into word streams. The performance of systems intended for this purpose can be influenced by many factors, including environment, vocabulary, speaker variability, etc. Technologies that involve speech processing and analysis cover areas such as voice recognition and speaker recognition and verification (Gupta, Bansal and Choudhary, 2018). According to Latif, Qadir, Qayyum, Usama and Younis (2021), automatic voice recognition consists of the following steps:

1. Acquisition: digitally obtaining the audio signal.
2. Pre-processing: noise suppression and silence removal.
3. Feature extraction: extract characteristics that describe the voice signal or represent it in a compact way. It is common to divide features into two types – linguistic and acoustic. The extraction procedures for these two types of characteristics are significantly different and performance depends on the type of problem to be solved.
4. Speech recognition: the use machine learning and deep learning algorithms (convolutional neural networks, recurrent neural networks and generative models) to process the speech to a written format.

Studies in other health care scenarios have used speech-to-text tools successfully, including in psychology studies in which verbal responses can be the most important outcomes (Ziman, Heusser, Fitzpatrick, Field and Manning, 2018). In this study, automated transcription proved to be equivalent to manual transcription, saving time and the need for human transcribers. Studies in radiology have also shown a great reduction in the time to deliver reports from 15.7 hours to 4.7 hours (Callaway, Sweet, Siegel, Reiser and Beall, 2002), as well as studies in the area of pathology in which the time to release reports was reduced from 4 to 3 days and the same-day report release rates increased from 22% to 37% (Singh and Pal, 2011).

Speech recognition technologies have already been subjected to a systematic review evaluating the use of these tools in different health areas (Johnson, Lapkin, Long, Sanchez, Suominen, Basilakis and Dawson, 2014). This study demonstrated benefits of this technology and that these benefits are directly related to the size of the transcription task (volume of information). Some of the disadvantages presented were the time to adapt to this tool, which generates unnecessary expectations on the part of employees who will use this type of solution.

On the other hand, an important assessment of expectations and experiences with voice recognition systems has already been carried out by Alapetite, Andersen and Hertzum (2009) with interesting results. This study used a questionnaire at the time of training the health professional and a questionnaire after 4 months or more using the

technology. A third (33%) of physicians approved its use, a third (31%) did not approve, and the remainder (34%) had neutral opinions. The authors' impression was that physicians' dissatisfaction was related to transcription errors and the time needed to make corrections to the information. Another interesting fact from this study was the perception of 94% of physicians that their time recording information has increased and 62% of them believe that the quality of medical records has worsened.

The study by Quiroz et al. (2019) identifies and discusses the main challenges associated with the development of automated speech-based documentation in clinical settings. In this study, issues such as the difficulties in obtaining medical data in sufficient volume to train machine learning algorithms, the difficulty in understanding the meaning of sentences according to the context, the complexity of the medical vocabulary and the non-linear progression are highlighted of subjects during the conversation.

A meta-analysis of 122 studies on the use of voice recognition software was performed by Blackley, Huynh, Wang, Korach and Zhou (2019). Most studies that evaluated this technology in emergency and radiology departments were included and found that studies in this area basically focus on three outcomes: productivity and time spent on medical records, errors in transcribing records and comparison with others transcription methods. In terms of time, the results were variable, with some studies showing a reduction of up to 19% of time compared to alternative methods of recording and 90% in relation to typing on the keyboard. All 19 studies that evaluated time to delivery of results demonstrated a significant improvement in this outcome, some with a 90% reduction in time (Zick and Olsen, 2001; Prevedello, Ledbetter, Farkas and Khorasani, 2014). This study demonstrated that the main concerns with the use of these recording tools are errors in medical records, which could potentially harm patients.

The above mentioned meta-analysis shows the predominance in the literature of research about speech-to-text only solutions for medical documentation. Herein we perform a systematic literature review on applications for speech recognition and simultaneous recording of patient-doctor interactions in a structured fashion.

4. Materials and Methods

This study aimed to established the current model and status in technologies for automatic transcription of health care professionals interactions with patients. This systematic review was conducted according to previously published guidelines (Biolchini, Mian, Natali and Travassos, 2005; Kitchenham and Charters, 2007; Popay, Roberts, Sowden, Petticrew, Arai, Rodgers, Britten, Roen, Duffy et al., 2006; Roehrs, Da Costa, da Rosa Righi and De Oliveira, 2017). The following steps describe the process:

1. Designing research questions
2. Designing a search strategy (Search term generation)
3. Development of a study selection criteria, procedure and quality assessment checklist
4. Creating a data extraction strategy
5. Selection of Databases (Sources of information)
6. Search term tests in all the Databases and search strategy improvement
7. Database search
8. Study selection
9. Quality evaluation of the selected studies
10. Data Extraction

4.1. Designing research questions

The research questions enumerated here were used to guide this investigation. They were divided into general and specific questions. The general questions aim at models for solving the problem of automatic transcription, results of applying the proposed solution and challenges faced by researchers in this area. Specific questions are related to the architecture of the solution; the machine learning methods utilized by specific models such as reinforcement learning, supervised learning and convolutional neural networks; and the results of the testing of the proposed solution. With specific questions we are also concerned with practical difficulties such as solutions to the problem of medical lexicon when translating symptoms to medical terms, and other practical problems such as solutions to the problem of room noises that could affect the transcription process. The questions were also used to elaborate the search terms, the data extraction strategy and to choose the correct databases for finding studies related to our objective.

General questions:

- GQ1: What are the intelligent models used to solve the problem of structured automatic documentation of the clinical history?
- GQ2: What are the challenges faced in creating an automatic structured documentation?
- GQ3: What problems could a structured automatic documentation help to solve?
- GQ4: How could structured automatic documentation improve medical care and doctor-patient interactions

Specific questions:

- SQ1: What are the specific machine learning methods and the architecture used to solve the structured automatic transcription of the clinical history?
- SQ2: How the patient symptoms were translated to medical terms when transcribing the clinical history?
- SQ3: How was the application tested (large scale tests such as in public health systems) and what were its results? Was there an improvement of care such as improved turnaround time and reduced documentation errors?
- SQ4: How was the acceptability of the program by its users, such as doctors and patients?
- SQ5: What are the commercial applications available for this purpose? Is there any data on the use of these applications?

4.2. Designing a search strategy (Search term generation)

The initial search terms were selected based on the general and specific questions. The scope of the search terms needed to be large enough so that all the questions were addressed and all the studies concerning this topic were found. Since this review was not focused in a complete and unedited transcription (strict speech-to-text) the search contained terms related to machine learning, artificial intelligence and natural language processing.

4.3. Development of a study selection criteria, procedure and quality assessment checklist)

In this phase we developed the selection criteria for the identification of primary studies, created the inclusion and exclusion criteria as well as a quality assessment checklist. We included only original articles that effectively proposed a computational solution to the problem of the automatic documentation of the interaction between health care professionals and patient. We excluded articles with the following criteria: scope reviews, commentary or letters, and strict speech-to-text studies. All the publications selected in the titles and abstract phase were used to find other publications that were not found by the initial search strategy.

4.4. Creating a data extraction strategy

Data extraction was based in the research questions proposed initially. We created a table containing the general and specific questions so that each one of the authors of this review could fill in the information directly extracted from the selected article. The information gathered in this phase was used to answer the research questions proposed initially and is the main focus of this work.

4.5. Selection of Databases (Sources of information)

This stage consisted in studying and listing databases to conduct the search. Since the subject of this review lies in the interface of technology and health related topics, a comprehensive list of Databases related to computer science, health and software engineering was contemplated.

4.6. Search term tests in all the Databases and search strategy improvement

We conducted an initial search through all databases utilizing the search terms initially proposed. We made several changes and refinements in the search terms during this process aiming in increasing the sensitivity and the number of articles contemplated as well as eliminating works not related to the topic in question, such as clinical trials of medical interventions to specific diseases. This process allows an improvement in the initial search for a broader and complete search.

4.7. Database search

The search was conducted through the selected databases on October 8th, 2021. The search databases utilized for this work were: Cochrane, IEEE, ACM, PubMed, Science Direct, Scopus and Web of Science. The search terms were similar but the search strategy needed to be different for each database, due to the differences in the logical operators of the search tools. The search terms and the number of publications for each database are detailed in Tables 1 and 2.

Database	Search terms
PubMed	I. (“voice” OR “speech”) AND (recogni*) AND autom* AND docum* II. (“natural language processing” OR “Machine learning” OR “artificial intelligence”) AND “speech recognition” [All Fields] AND automat*[All Fields] AND transcrip* [All Fields]
Web of Science (Topic)	(medic* OR health*) AND (“speech recognition” OR “voice recognition”) AND (natural language processing OR machine learning OR artificial intelligence) Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI
Association for Computing Machinery (ACM) Digital Library	I. Abstracts search: medic* AND speech AND recog* NOT (therapy OR acoustic OR emotion OR brain OR implant OR disorder OR messag* OR signal* OR imag* OR disabled OR translate) II. Abstracts search: speech AND recog* AND healthcare NOT (therapy OR acoustic OR emotion OR brain OR implant OR disorder OR messag* OR signal* OR imag* OR disabled OR translate) III. Abstracts search: medic* AND dicta* NOT (therapy OR acoustic OR emotion OR brain OR implant OR disorder OR messag* OR signal* OR imag* OR disabled OR translate) IV. Abstracts search: medic* AND speech AND dicta* NOT (therapy OR acoustic OR emotion OR brain OR implant OR disorder OR messag* OR signal* OR imag* OR disabled OR translate) V. Abstracts search: (medic* OR healthcare) AND (“natural language processing” OR “machine learning” OR “artificial intelligence”) AND (“speech recognition” OR “voice recognition”) NOT (therapy OR acoustic OR emotion OR brain OR implant OR disorder OR messag* OR signal* OR imag* OR disabled OR translate)
IEEE Xplore	I. (“voice recognition” or “speech recognition”) AND (medic* OR healthcare) AND (“natural language processing” OR “machine learning” OR “artificial intelligence”) II. dictation AND medic* III. (“speech recognition" OR “voice recognition”) AND software AND (medic* OR healthcare) NOT (therapy OR acoustic OR brain OR implant OR disorder OR imag* OR disabled OR translate OR training)
ScienceDirect (Elsevier)	“natural language processing” AND healthcare AND (speech recognition OR voice recognition) AND (“artificial intelligence” OR “machine learning”) AND NOT (therapy OR implant)
Cochrane Database of Systematic Reviews	I. (speech recognition OR voice recognition) AND software II. speech recognition OR voice recognition
Scopus	((TITLE-ABS-KEY(“speech recognition”) OR TITLE-ABS-KEY(“voice recognition”)) AND (TITLE-ABS-KEY(“medical”) OR TITLE-ABS-KEY(“health”)) AND (TITLE-ABS-KEY(“natural language processing”) OR TITLE-ABS-KEY(“machine learning”) OR TITLE-ABS-KEY(“artificial intelligence”)) AND NOT (TITLE-ABS-KEY(“hearing”) OR TITLE-ABS-KEY(“therapy”))

Table 1: Search terms for each database

Database	Number of Articles Yielded
PubMed	89
Web of Science (Topic)	256
Association for Computing Machinery (ACM) Digital Library	341
IEEE Xplore	227
ScienceDirect (Elsevier)	569
Cochrane Database of Systematic Reviews	14
Scopus	626

Table 2
Number of publications for each database

ID	Question
QE1	Does the article clearly show the purpose of the research?
QE2	Does the article adequately describe the literature review, background, or context?
QE3	Does the article have an architecture proposal or research methodology described?
QE4	Does the article present the related work with regard to the main contribution?
QE5	Does the article have research results?
QE6	Does the article present a conclusion related to the research objectives?
QE7	Does the article recommend future works, improvements, or further studies?

Table 3
Quality evaluation questions

4.8. Study selection

Titles and abstract potentially eligible were screened independently by FSF and FKA. Disagreement in the initial selection was solved through discussion. The full-text evaluation for inclusion in the systematic review and data extraction were conducted for the full texts in the same manner by the same authors using a standardized form.

4.9. Quality evaluation of the selected studies

Quality evaluation is an important part of systematic reviews, since it gives weight to the information gathered. To assess the quality and strength of the selected articles we used the set of questions proposed by (Roehrs et al., 2017). This questions evaluate the purpose of the research, background and context, literature review, related work, proposed architecture and methodology, results, and the relation of the conclusion to the research objectives. Quality evaluation was executed by FSF and FKA independently. The quality evaluation questions are shown in Table 3.

4.10. Data Extraction

In this phase two authors (FSF and FKA) read the selected studies and completed the data extraction table. All the disagreements were solved through discussion and a third author (JCSL) helped solving discrepancies. A structured table with global and specific questions was constructed for data extraction. Data gathered through this process was discussed by three authors (FSF, FKA and JCSL) and synthesized as the core information of this review.

5. Results

The search resulted in 1995 titles with 113 duplicates. In the titles and abstracts screening 1825 articles were excluded and 55 entered the full text evaluation for eligibility. In the full text evaluation stage 6 studies met the inclusion criteria and two new studies not found by our search strategy were found as references from other articles (Klann and Szolovits, 2009; Khattak, Jeblee, Crampton, Mamdani and Rudzicz, 2019), which resulted in 8 studies with data to be extracted. Studies that dealt only with speech-to-text technology was the principal reason for exclusion with 26 studies in this category. We found 6 studies that were classified as scope reviews, 1 study classified as commentary or letter, and 1 systematic review. The remaining studies were excluded for other reasons (evaluation of acceptance of clinical

documentation by voice, automatic creation of subtitles for the hearing impaired, text-only information extraction, development of a program to help people with speech handicaps...). The flowchart is presented in Figure 3 and the result of the data extraction in Tables 4 and 5.

Study ID	**GQ1: What are the intelligent models used to solve the problem of structured automatic documentation of the clinical history?**	**GQ2: What are the challenges faced in creating an automatic structured documentation?**	**GQ3: What problems could a structured automatic documentation help to solve?**	**GQ4: How could structured automatic documentation improve medical care and doctor-patient interactions?**
ahamed2021 (Ahamed, Weiler, Boden, Januschowski, Stennes, McCrae, Bock, Rawein, Petris, Foth et al., 2021)	Convolutional neural networks; Long Short-Term Memory neural networks (LSTM)	High volume of data to be processed by the ASR module (49.6% of the speech recorded is relevant for documentation in medical examinations of patients receiving intravitreal injections); Filtering conversation data for segments relevant for documentation	Reduce time consumed in documentation during medical care	Reduction of the repetitve process of documentation; Increase time for doctor-patient conversation
finley2018 (Finley, Edwards, Robinson, Sadoughi, Fone, Miller, Suendermann-Oeft, Brenndoerfer and Axtmann, 2018)	Speaker diarization; Automatic speech recognition; Knowledge extraction; Natural language generation	Speaker indexing (who spoke when); Extracting information from spontaneuous conversational speech	Reduce work time in tasks related to EMR; No necessity of human scribe (high turnover, training time, cost)	Less time with EMR and more time dedicated to doctor-patient interactions
guisi2014 (Ni, Shi and Mahajan, 2014)	Speech recognition component; Keyword search; Assistant Central Component	Distinguishing different speakers; Producing an output compatible with the eletronic medical record	Lack of standardized approach to patient evaluation and subotimal history collection by doctors and trainees; Time constraints; Language and cultural differences between patient and physician	Augment patient physician-interaction and enhance patient satisfaction; Ensure completeness of history taking/physcial examination; Reduce costs of care and manual effort; Optmize residents of all medical and surgical fields
khattak2019 (Khattak et al., 2019)	Utterance type classification; Time expresssion identification; Medical entity identification; Attribute classification; Primary diagnosis classification	Classifying pertinent entities for a diagnosis; Limitation of medical terms from the reference lists; Accounting for similar medical terms and spelling variations	Reducing time to enter information in EMR; Improve data consistency and completeness in EMRs and help prospective data analysis	Doctors could spend more time talking to patients; Some computer programs could be applied to the documentation to predict information such as disease probabilty and mortality
klann2009 (Klann and Szolovits, 2009)	Automated speech recognition; Medical natural language processing	Integration of multiple program modules; Difficulties in capturing the audio from the interview (information from health professional and patient); Necessity of a large database of conversations to train the program; Need to develop a voice-controlled program to correct what was already transcribed	Solve defficiencies in medical records that often fail to include critical information	Improve aspects of translational medicine such as the interaction of clinical data with genome information; Help researches to identify useful informations
maas2021 (Maas, Kisjes, Hashemi, Heijmans, Dalpiaz, Dulmen and Brinkkemper, 2021)	Not clearly described	Unconstrained text dialogue for analysis; Need for a robust architecture; Balance between required expressiveness and computational demands in constructing a formal representation of the transcriptions; Differences between hospitals on terminoloy and procedures	High medical workload due to the large amount of documentation; Necessity of maintaining quality of patient data	Improve administrative efficiency and personal engagement in health care; Reduce time necessary to record and maintain appropriate EMR
wenceslao2019 (Wenceslao and Estuar, 2019)	Speech recognition module; Natural language processing module; Summarization module	Internet connection; Ensuring patient data privacy; System usability; Environment noises, dificulties in word and context recognition and transcription delays	Reduce doctor fatigue and consultation time; Improve usability of most EMRs; Possibility of recording the consultation for legal uses (with patient authorization)	Increase doctor-patient interactions; Improve usability of health applications for better value, maintaining patient safety, and improving efficiency

woo2021 (Woo, Mishra, Lin, Kar, Deas, Linduff, Niu, Yang, McClendon, Smith et al., 2021)	Noise-resilient ASR; Multi-style training; Customized lexicon; Speech-enhancement	Errors in machine transcriptions in noisy enviroments (pre-hospital locations); Detecting medical and military terms used by the medical professionals in an emergency situation	Incomplete or failed documentation of prehospital care in emergency situations; Medical errors related to insufficient documentation or communication in the transition from prehospital to hospital care	Reduce clinical errors (duplicate administration of the same medication); Hand-free documentation in the field during prehospital care; Transmit complete and accurate informations from the point of injury to the field hospital

Table 4: Global questions

Study ID	**SQ1: What are the specific machine learning methods and the architecture used to solve the structured automatic transcription of the clinical history?**	**SQ2: How the patient symptoms were translated to medical terms when transcribing the clinical history?**	**SQ3: How was the application tested (large scale tests such as in public health systems) and what were its results? Was there an improvement of care?**	**SQ4: How was the acceptability of the program by its users, such as doctors and patients?**
ahamed2021 Ahamed et al. (2021)	Convolutional neural networks with 2-10 convolutional layers, 2-10 dense layers, 512 neurons per layer, Log-mel spectrogram for extracting features; Recurrent neural networks (Long Short-Term Memory networks and Bi-directional LSTM) with 1-4 LSTM layers, 1-4 dense layers, Up to 1024 neurons per layer. Use of Python, Keras deep learning library and Tensorflow. The audio was first down-sampled to 20 kHz and then segmented in utterances utilizing an engine devoloped at the Fraunhofer IDMT	Not described	Tested in 69 routine medical examinations of patients receiving intra-vitreal injections recorded during follow-up visits in the Eye Clinic Sulzbach (Germany); The information in conversations was manually classified in relevant or irrelevant; Convolutional neural network showed the highest validation accuracy; Dataset B (separated by gender): Accuracy: 0.9718, Precision: 0.9472, Recall: 0.9839, AUC: 0.996; Dataset C (separated by accent): Accuracy: 0.9708, Precision: 0.9545, Recall: 0.9927, AUC: 0.9961; Dataset D (separated by selected speaker): Accuracy: 0.9754, Precision: 0.9658, Recall: 0.9857, AUC: 0.987	Not described
finley2018 Finley et al. (2018)	Speaker diarization: "top-down approach", a Hidden Markov, Gaussian mixture model to represent the likely audio features and timing characteristics of dialogs; A modified expectation maximization algorithm at decoding time was used to learn the current speaker and background silence characteristics in real time; ASR: a neural network trained to predict context-sensitive phones from the audio features; Knowledge extraction: classification of turns in the conversation based upon the information they likely contain using hierarchical recurrent neural networks	Information extraction strategies: rule-based processing to identify predictable elements; knowledge-based strategies (semantic overlap with dictionary definitions, for variable concepts such as symptom descriptions); and fully supervised machine learning approaches for complex tasks; Sentence templates in a sentence bank to structure the final report; Natural language generation: data-driven template + finite-state "grammar" of report structure; Sentence bank with a sentence template; Separate natural language generation models for different hospitals and specialties	The application is still in development	The application was not tested

guisi2014 Ni et al. (2014)	Server-based ASR: google speech-to-text API (input to the acoustic and grammar model); Assistant Central Component: keyword search (methodology not described); Clinical Framework: possibility to customize keywords and supplemental information (changes will be stored in HTML form as plist file and connected to local database or cloud database)	Not described	A single case study was reported with 82.17% accuracy at word recognition and increased questions asked by the physician from three to eight	The application was not tested
khattak2019 Khattak et al. (2019)	Every utterance in the dialogue is labeled as a question, statement etc. A two-layer bidirectional gated recurrent unit neural network, implemented in PyTorch was used; Each word is represented as a 200-dimensional vector using the freely available Wikipedia-PubMed word embedding model; Medical entity identification uses lexicon look-up using terms from BioPortal, Consumer Health Vocabulary, SNOMED-CT, and RxNorm. Attribute classification: a support vector machine trained with stochastic gradient descent to classify modality (actual, negative, possible) and pertinence (medical entity). Primary diagnosis classification: tf-idf on the cleaned text of each dyad and logistic regression, SVMs, and random forest models.	Lexicon look-up from BioPortal, Consumer Health Vocabulary, SNOMED-CT, and RxNorm	Data: 800 audio patient-clinician dialogues and transcripts (Verilogue INC). Results: Utterance type classification: F1 score: 0.71; Medical entity identification: F1 score of 0.63 and 0.55 Krippendorff's alpha; Attribute classification: F1 score of 0.77 for modality and 0.62 for pertinence; Primary diagnosis classification: F1 scores (Linear SVM): Influenza .93±.04, ADHD .83±.05, COPD .68±.14, Osteoporosis .78±.04, Type II diabetes .76±.07, Depression .71±.08, and Other .76±.05.	Not described
klann2009 Klann and Szolovits (2009)	Dragon's Naturally Speaking ASR system (from Nuance); Medical natural language processing: Category and Relationship Extractor (CaRE) implemented in Java and Perl invoking Support Vector Machine (SVM), the Brill tagger, Link Grammar Parser and UMLS metathesaurus; GATE, from Sheffield University, UK for integration of the different components;	Software Development Kit from Dragon's Naturally Speaking + Medical Edition (both from Nuance) for better accuracy on medical terminology	No large scale tests	The application was not tested
maas2021 Maas et al. (2021)	Large unimodal analyzers (audio, video, domotics) with predefined input and output sets; Google cloud speech-to-text; Database with data structure correspondence with medical consultation timeline; Extraction of information (triples) to populate the PMG by linguistic tools, ontology development and triples stored and managed with StarDog (linguistic annotation using Python Frog); Vídeo analysis: OpenCV and YOLO libraries; Medical guidelines modeled in PROforma;	Corpus of medical background knowledge: Medical ontologies (SNOMED, ICD-10, LOINC) and Medical knowledge graphs (Drugbank, SIDER, AERS); Creation of patient medical graphs (anatomy, symptom, observation, diagnosis and treatment); Capture medical dialogue through a library of linguistic patterns with placeholders to be filled by speech tagging; Report generation: template of sentences generated by the patient medical graph	No large scale tests	The application was not tested

wenceslao2019 Wenceslao and Estuar (2019)	IBM Watson Speech to Text and the HTML5 Web Speech API (using Google Chrome); Natural language processing module: Clinical Text Analysis and Knowledge Extraction System (cTAKES) software for annotation; Summarization module: A client-server implementation of cTAKES was used by an application that acts as a Representational State Transfer (REST) service wrapper. MyLifeEMR receives JSON objects returned by this REST service that are mapped to Sign Symptom Mention, Measurement Annotation, Disease Disorder Mention, Medication Mention, and Procedure Mention	cTAKES cross-checked entities with dictionaries of the Unified Medical Language System	Tested by one doctor in one session; System usability scale was used to grade the performance of the prototype	The prototype scored 45 in a scale from 0-100 (the prototype failed the usability test filled by the doctor)
woo2021 Woo et al. (2021)	Software Architecture: Open-source ASR platform (Kaldi); Natural language processing module: generating bookmarks for preliminary documentation; Time delay neural network (TDNN)(each layer with 1536 nodes): 1 input layer; 11 TDNN layers; 1 linear output layer; A dictionary with domain-specific words and their corresponding phonemes was generated; MetaMap 2018 was used as a post natural language processing module for medical information extraction	A new customized lexicon was trained from medical and military terms used in battlefield-related injuries and medical evacuation; Information was collected from the staff taking care of the patient in a combat or rescue situation	6 focus groups with 26 individuals (emergency medical services, transport nurses, and emergency department physicians): Semistructured interview; 21 simulation drills over 3 days; Short debriefing interviews to gathered feedback; Postsimulastion focus groups; ASR trained with noisy audio data set; The program was tested in three clinical loosely structured simulations: 27 complete patient cases spanning from field to field hospital were simulated and collected (5.05 hours of audio recordings); The program was compared to human medical transcriptionists who listened to the audio recording of all simulations and to the Dragon Medical Practice Edition 4 (DMPE 4) software; Results: Baseline + multi-style training + updated language model + speech enhancement: Medical word error rate: 46.3% and F1 score of 0.781	Participants emphasized the need for device flexibility, sturdiness and lightweight; Participants recommended training with the device before using it in real clinical scenarios and emphazied the need to talk loud during its use; Short notes, recording, and video and photos from the injuries could help the application increase in value as well as help the site to which the patient is being trasported

Table 5: Specific questions

5.1. GQ1: What are the intelligent models used to solve the problem of structured automatic documentation of the clinical history?

The models that were developed by the researchers have been described with very uneven degrees of completeness and detail. In one instance, as in the paper by Maas et al. (2021) how the intelligent system worked was not described at all. In Ahamed et al. (2021) al it was briefly reported that an artificial neural network was programmed in Python using the Keras deep learning library with a Tensorflow backend. Different architectures were tested, and in the end the Convolutional Neural Networks (CNNs) were found to be more successful than Long Short-Term Memory (LSTM) networks. Finley et al. (2018) on the other hand described a processing pipeline in four major stages: speaker diarization, automatic speech recognition, knowledge extraction, and natural language generation, with each step being explained in details in one section of the paper.

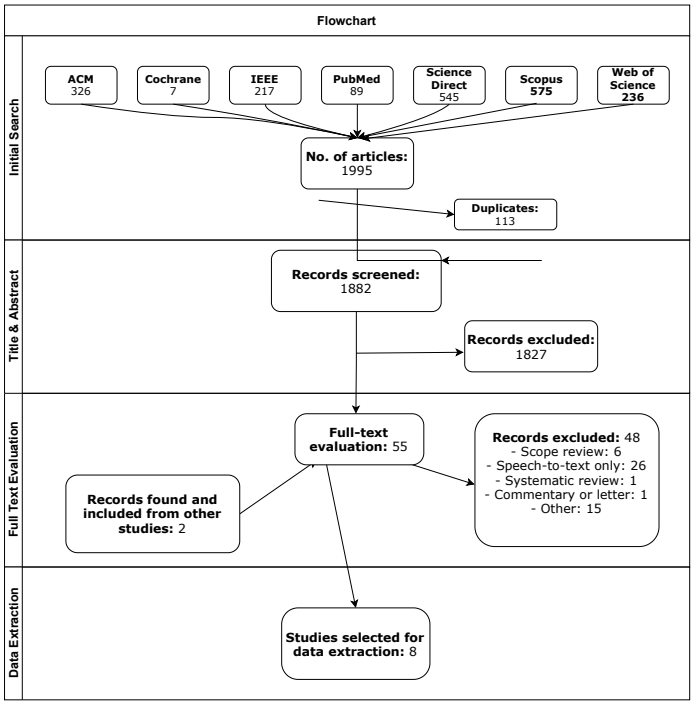

Figure 3: Flowchart.

In a similar fashion, Ni et al. (2014) reported an iOS platform based speech recognition system in which voice content is converted to text, keywords are searched and supplementary questions are suggested. The speech recognition component, keyword search and assistant central component were accordingly described. The system developed by Khattak et al. (2019) was named AutoScribe, a system in several modules for automatically extracting pertinent medical information from medical interactions. The system classifies phrases for relevance and extracts medically relevant information through natural language processing.

Klann and Szolovits (2009) thoroughly described a proof-of-concept system in Java that permits a lash-up of Dragon's Naturally Speaking (DNS), an automated speech recognition system, with an natural language processing system called Category and Relationship Extractor (CaRE) and General Architecture for Text Engineering (GATE), the comprehensive component integrate the different components of this task. In Wenceslao and Estuar (2019) two web-based services were developed to recognize speech, process it and structure it in a SOAP (subjective, objective, assessment, plan) model. Finally, an application was created for automatic speech recognition and structured documentation of emergency medical assistance in combat or rescue environments, so this system was worked to be noise resilient and trained in the specific lexicon of this type of situation (Woo et al., 2021).

5.2. GQ2: What are the challenges faced in creating an automatic structured documentation?

The most frequent challenges faced by researchers in developing such applications were related to singularities of the medical interview, like understanding medical terminology, once medical vocabulary is very complex, processing high volumes of information and separating parts that are relevant for registration from those that are not. Medical terms may be named by acronyms or only by the initials, and terminology and interview structure may be very different from one hospital to another and even between different clinics in the same institution. At least one author, admitted that to train a program in that scenario would require large sets of data (Klann and Szolovits, 2009).

A large amount of variables that were not clearly mentioned in the studies can interfere in the success of using solutions for automatic structured documentation. Among them, must be highlighted concerns about patient privacy and safety, as well as transcription errors and the time needed to make corrections to the information, that often cause fatigue and discontent. In addition, accents can vary significantly, Internet connection may be low quality or may not

be available, microphones may fail and users (doctors) may be poorly trained in the use of these systems. However, quality assurance procedures, auditing and comprehensive training have potential to help improve the quality of these systems.

Many other challenges were difficulties that are common to any other voice recognition application: voice detection and transcription, mostly in noisy environments, speaker indexing and distinguishing different speakers. Some singularities of the medical interview might make these problems particularly challenging in the health care context. First, more than one person in the room can be sometimes the rule and not the exception, as many patients have family members together when visiting a doctor. Depending on the patient, the presence of an accompanying person can be a necessity, as the information acquired from the patient himself could be totally unreliable. It has not not yet been described how the system would adjust and correct itself in relation to different sources of information during the interview. Second, interruptions and noise are common in any medical practice, and they would make the transcription particularly difficult in emergency rooms and inpatient units. One of the papers paid a lot of attention to this issue because the register was executed in pre-hospital emergency situations (Woo et al., 2021). Third, we could possibly say that one of biggest challenges that applications designed for automatic registration of the medical encounter will face is how to account for the unspoken parts of the evaluation. Even though this point was not extensively explored in the papers, a substantial part of the patient assessment by a physician comes from the physical examination, inspection, speech analysis and physical mobility evaluation. How could these aspects be integrated in the record without requiring the physician to manually type the data is not yet known.

5.3. GQ3: What problems could a structured automatic documentation help to solve?

The most frequent aspect that researchers thought automatic documentation would help solving was time expended typing and working in the EMR. This is a problem that has been related to burnout, increased physician working hours, reduced time with patients and less patient satisfaction during the encounter as has already been explained above. Solving this issue would be the main benefit of the automatic documentation. One previous solution to this problem was a human scribe, but as Finley et al. (2018) has mentioned in his paper, a computer scribe would be less costly and would keep the interview more private for both the patient and the health professional. The other main application would be to improve the medical records by itself by ensuring completeness, thoroughness and avoiding mistakes, omissions and copying of previous registers. That would not only be a gain for patient safety and quality of care, but also would be helpful for medical research as well as useful in cases of litigation.

In addition, automatic documentation systems could interact with other solutions, such as algorithms that might suggest doctors to ask additional questions during the appointment or order additional tests. The suggestions of these algorithms could be classified by physicians as sufficient, partially adequate or inadequate. There could also be options for the physician to configure the system with the type of suggestion he would like to receive, as well as other desirable behaviors in the automatic documentation system. These features increase the physician's autonomy regarding the use of the system and, in this way, can contribute to improve their satisfaction and motivation.

5.4. GQ4: How could structured automatic documentation improve medical care and doctor-patient interactions?

The first benefit of automatic documentation would be to increase the time doctors expend with their patients and to offer a more humane and empathetic interaction, as have almost all the articles selected for this review acknowledged. Ideally, a system for automatic documentation generation is easy to use, works with little user interference, requiring little time to correct information. So, doctors could spend more time interacting with patients instead of looking at the computer screen and this can contribute to the improvement of care and increase the satisfaction of both parties.

Cost was another important topic, as the automatic register could increase physician efficiency and reduce administrative human resources. In places where human scribes are used, these professionals could be reassigned for other tasks. Billing and coding today require a lot of time and attention from both physicians and administrative personnel:these tasks could be automatically performed by a computer, increasing efficiency and reducing burnout.

Other authors like Khattak et al. (2019) and Klann and Szolovits (2009) also mentioned the possibility of integrating EMRs with large clinical and genomics databases to make predictions of mortality, time to return to the hospital after discharge, drug toxicity and risk of cancer. One of the publications selected (Woo et al., 2021) was specifically interested in the possibilities of this application for rescue and emergency missions, when medical register is difficult and, as in many instances, impossible. Automatic register would allow the rescue team to have their full attention in saving the victim´s life in extreme circumstances.

The use of automatic documentation solutions has the potential to contribute to obtaining documents that are richer in detail and with better quality, which can result in more correct diagnoses. In many countries physicians don´t have a culture of producing adequate registers or they simply don´t have enough time to do it, what makes it very difficult to find complete and reliable documentation in retrospect. In addition, reducing the stress of health professionals can also contribute to the improvement of care and allow more time to be invested in more rewarding tasks. Finally, good quality documents can be useful in research. Overall it is expected that automatic registration would ensure more accurate and efficient documentation of all the relevant aspects of health care.

5.5. SQ1: What are the specific machine learning methods and the architecture used to solve the structured automatic transcription of the clinical history?

The study from Ahamed et al. (2021) compared convolutional neural networks (CNN) with recurrent neural networks (RNN) (Long Short-Term Memory networks and Bi-directional LSTM). The CNNs were constructed with 1-10 convolutional layers, 2-10 dense layers with 512 neurons per layer and the RNNs were programmed with 1-4 LSTM layers, 1-4 dense layers with up to 1024 neurons per layer. This software was developed utilizing Python with the Keras deep learning library and TensorFlow and the features were extracted using Log-mel spectrogram.

The study by Finley et al. (2018) utilized an architecture involving speaker diarization, automatic-speech recognition, knowledge extraction and natural language generation. Speaker diarization concerns the problem of identifying two different speakers and was solved utilizing through a representation of audio features and transition characteristics of dialogues through a Hidden Markov, Gaussian mixture model. To identify the current speaker a modified expectation maximization algorithm was used at decoding time and this data was then processed by a neural network composed of an acoustic model and a language model trained to predict context-sensitive phones from the audio features. Knowledge extraction was solved through different methodologies such as recurrent neural networks and supervised machine learning.

The study by Ni et al. (2014) presented an interesting software to help training physicians to collect better clinical history with suggestions of complementary questions. This system utilized the Google speech-to-text API as a server-based ASR solution and created a customizable database so that the user could alter the information that should be gathered for a specific disease. One of the studies utilized Dragon's Naturally Speaking ASR system, which is a commercially available ASR software from Nuance and its data was used as input to a listening framework implemented in Java and run on Windows (Klann and Szolovits, 2009). A medical natural language processing was implemented in Java and Perl using Support Vector Machine, brill tagger, link grammar Parser and the Unified Medical Language System metathesaurus to identify semantic types of words. All of these components were integrated with the General Architecture for Text Engineering, an open source software for text processing (Cunningham, Tablan, Roberts and Bontcheva, 2013).

Maas et al. (2021) created a software using large unimodal analyzers to capture audio, video and domotics with predefined input and output sets and used Google cloud speech-to-text API as a processing tool. A knowledge graph, entitled patient medical graph, was constructed based on subgraphs of anatomy, symptom, observation, diagnosis and treatment. These subgraphs were composed of ontologies concerning each of its domains and were mapped as placeholders for the information gathered in the consultation. To populate these placeholders the information was stored in triples, managed with StarDog and annotated with Python Frog. The back and front-end were written in C# and the software run on Universal Windows Platform.

The study by Woo et al. (2021) processed the audio with a deep neural network model based on Speech Enhancement Generative Adversarial Network (SEGAN) to minimize background noise and used an Kaldi as an open-source ASR platform. A language model using probability distribution was constructed to infer words based on context using probability distribution over sequence of word and a processing model generated bookmarks for documentation. A time delay neural network (TDNN) with 1536 nodes in each layer was designed with 1 input layer, 11 TDNN layers and 1 linear output layer. This TDNN was trained based on the aligned frames and senones obtained through a Gaussian mixture model–hidden Markov model. Terms from battlefield-related injuries and medical evaluation were used to train a customized lexicon and the medical information was extracted using MetaMap 2018.

In the study by Khattak et al. (2019) the utterances in the dialogue were labeled as questions, statements etc. A two-layer bidirectional gated recurrent unit neural network, implemented in PyTorch was used. Each word is represented as a 200-dimensional vector using the freely available Wikipedia-PubMed word embedding model. Named entity recognition and classification is a text analysis technique based on NLP to automatically pull out specific data from unstructured text, and classifies it according to predefined categories (named entities), the words or phrases that

represent a noun. In this study, the medical entity identification uses lexicon look-up using terms from BioPortal, Consumer Health Vocabulary, SNOMED-CT, and RxNorm. Attribute classification is performed by a support vector machine (SVM) trained with stochastic gradient descent to classify modality (actual, negative, possible) and pertinence (medical entity). Primary diagnosis classification uses tf-idf (a statistical calculation adopted by Google's algorithm to measure which terms are most relevant to a topic) on the cleaned text of each dyad and logistic regression, SVMs, and random forest models.

In the study by Wenceslao and Estuar (2019) was used the IBM Watson Speech to Text and the HTML5 Web Speech API (using Google Chrome). In the NLP module cTAKES software was used for annotation. A client-server implementation of cTAKES was used by a GitHub application by GoTeamEpsilon that acts as a Representational State Transfer (REST) service wrapper accessed through a specified address. REST, in turn, is a set of architectural constraints for services. Regarding the summarization module, MyLifeEMR receives JSON objects (JSON is a lightweight data-interchange format) returned by the REST service are mapped to Sign Symptom Mention, Measurement Annotation, Disease Disorder Mention, Medication Mention, and Procedure Mention. In addition, blockchain technology is used for data safety.

5.6. SQ2: How the patient symptoms were translated to medical terms when transcribing the clinical history?

A document written by a physician containing the clinical history of a patient is very different from the real interview. It contains multiple medical terms used to describe with precision symptoms that could take too many words to be correctly reported. This is an important problem for the automatic documentation, since the application must change the words from the interview to adequate medical terms. Finley et al. (2018) used knowledge-based strategies such as semantic overlap with dictionary definitions, for symptom description. This structured representation allowed a scribe to evaluate the information extracted before the generation of the final report. After this process they used a sentence bank with templates, this templates were annotated to receive specific data types. After this sentences were filled with information they were clustered in a standardized format.

The application proposed by (Khattak et al., 2019) a lexicon look-up of terms from BioPortal, SNOMED-CT, Consumer Health Vocabulary, and RxNorm to identify entities like anatomical locations, signs and symptoms, diagnoses, and therapies. The software developed by Klann and Szolovits (2009) integrated in their program the Software Development Kit from Dragon's Naturally Speaking by Nuance as well as its Medical Edition for better accuracy on medical terminology. The study by Maas et al. (2021) utilized as medical background SNOMED, ICD-10 and LOINC and Drugbank, SIDER and AERS to build the knowledge graphs. The dialogue was converted to text and the information was tagged and used to fill a library of linguistic patterns. From that a report was generated using these constructions in a pre-specified structure"

The program proposed by Wenceslao and Estuar (2019) used the software cTAKES developed by the Mayo Clinic and cross-checks detected the input text extracted entities with the dictionary the Unified Medical Language System. Lastly (Woo et al., 2021) trained for their pre-hospital automatic documentation device a new customized lexicon from medical and military terms commonly used in this clinical scenario using the Carnegie Mellon University Sphinx Knowledge Base Tool (Walker, Lamere, Kwok, Raj, Singh, Gouvea, Wolf and Woelfel, 2004). This new lexicon was merged with the original dictionary and language models and this was used as the new lexicon (Woo et al., 2021).

5.7. SQ3: How was the application tested (large scale tests such as in public health systems) and what were its results? Was there an improvement of care such as improved turnaround time and reduced documentation errors?

None of the studies selected presented large scale tests in the health systems. Of the studies selected, five of them presented some kind of test of their applications (Ahamed et al., 2021,?; Wenceslao and Estuar, 2019; Woo et al., 2021; Ni et al., 2014; Khattak et al., 2019). The study by Ahamed et al. (2021) tested its module in classifying relevant and irrelevant information in the anamnesis. The test consisted in 69 recordings of routine medical examinations of follow-up visits of patients receiving intra-vitreal injections in the Eye Clinic Sulzbach in Sulzbach, Germany. The performance of the program was compared to a manual classification of the information and demonstrated a validation accuracy of 92.41%.

The application by Ni et al. (2014), that aimed in assisting the physician in improving the clinical history as well as the efficiency of the patient-physician communication was tested in a pre-determined single case scenario. In this test the study showed 82.17% accuracy at word recognition and important increase in questions asked by the physician from

three to eight. The AutoScribe proposed by Khattak et al. (2019) was tested on 800 audio patient-clinician dialogues and their transcriptions purchased from Verilogue INC and showed F1 scores of 0.71 for utterance type classification, 0.63 for medical entity identification, and 0.77 for attribute classification. For primary diagnosis classification it showed F1 scores of 0.93 for influenza, 0.83 for ADHD, 0.68 for COPD, 0.78 for osteoporosis, 0.76 for type II diabetes and 0.71 for depression.

The study by Woo et al. (2021) tested their prototype with 6 focus groups with a total of 26 individuals from different health care areas such as emergency medical services, transport nurses, and emergency department physicians). This test consisted of a semi-structured interview followed by simulation drills (21 sessions over 3 days). The sessions consisted of 27 complete patient cases spanning from field to field hospital and were recorded in approximately 5 hours of audio that were latter transcribed by human medical transcriptionists and by the Dragon Medical Practice Edition 4 software for comparison. In this comparison the best model proposed achieved a medical word error rate: 46.3% and F1 score of 0.781.

5.8. SQ4: How was the acceptability of the program by its users, such as doctors and patients?

Only two studies presented user tests results in live clinical scenarios (Wenceslao and Estuar, 2019; Woo et al., 2021). The study by Wenceslao and Estuar (2019) tested the usability of the prototype with one doctor in a clinical simulation. The user was described as having more than 5 years of experience and deep knowledge of the local EMR. After the test the participant was asked to fill an enhanced system usability scale that grades the product/prototype in a scale from 0 to 100 with an average of 68, above which the product is classified as being generally usable (Brooke et al., 1996). The prototype failed the usability test with a result of 45 in a scale from 0 to 100, a clear result that prompt researchers in this field that usability is a top priority for this kind of application. As a side note the study also compared two different speech-to-text API, the Google-powered HTML5 Web Speech API and the IBM Watson. In this comparison the Google API was considered a better solution since it is compatible with multiple languages and showed very low latency of transcription. On the other hand the IBM Watson showed a lag time of more than a minute to process after the 1 to 2 minute test case was finished.

The participants testing the device from the study by Woo et al. (2021) emphasized the need for device flexibility, sturdiness and lightweight since it was developed for pre-hospital care documentation. Training with the device before its use in real clinical scenarios and the need to talk loud during its use were some of the other recommendations. The participants also suggested that the possibility of including short notes, recording the clinical scenario, and video and photos from the injuries could help the application increase its value as well as help the site to which the patient is being transported with valuable information.

5.9. SQ5: What are the commercial applications available for this purpose? Is there any data on the use of these applications?

None of the studies selected presented a commercial application or large scale data on its use. For this reason, SQ5 was suppressed in table 5. However, there are significant commercial applications like the Dragon Ambient Experience from Nuance and a similar product from Augmedix. Other companies such as Sopris Health, Orbita, Sensely, Suki, Notable and CareVoice have products with this purpose and must be cited, however no study was found testing this products or describing their architecture (Topol, 2019). This is the case where we may be experiencing the fact that the innovation and scientific breakthrough of the proposed application is closely related with commercial gains and there is no interest in academic publications.

Dragon Medical One, the product developed by Nuance, has excellent reviews and seem to solve most of the problems encountered in our research. This application can interact with the EHR and search UpToDate and MDCalc for clinical evidence. In the same direction the Augmedix application can also interact with the EHR and is described, in its website, in a more general way to be a combination of automated speech recognition, natural language processing and machine learning to generate medical notes.

5.10. Quality evaluation

All the studies selected for this systematic review showed a clear purpose of research, even though some of the studies did not described an adequate literature review or scientific context for their application/program. All of the studies demonstrated some kind of architecture for their work, even though a more thorough description could increase the reproducibility of the works proposed as in the study by Finley et al. (2018). Three of the studies lacked test results (Finley et al., 2018; Klann and Szolovits, 2009; Maas et al., 2021) and all of them recommended future work since

Study ID	QE1	QE2	QE3	QE4	QE5	QE6	QE7
ahamed2021 Ahamed et al. (2021)	Yes	No	Yes	Yes	Yes	Yes	Yes
finley2018 Finley et al. (2018)	Yes	No	Yes	Yes	No	No	Yes
guisi2014 Ni et al. (2014)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
khattak2019 Khattak et al. (2019)	Yes	No	Yes	No	Yes	Yes	Yes
klann2009 Klann and Szolovits (2009)	Yes	Yes	Yes	Yes	No	Yes	Yes
maas2021 Maas et al. (2021)	Yes	Yes	Yes	Yes	No	Yes	Yes
wenceslao2019 Wenceslao and Estuar (2019)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
woo2021 Woo et al. (2021)	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 6
Quality evaluation answers

this is not a complete conquered area. Also important to note that none of the articles had a commercially available product at the time of the publication. The Table 6 shows the quality evaluation answers.

6. Challenges and Future Directions

We have performed a systematic literature review on intelligent solutions for automatic speech recognition with automatic documentation during a medical interview. The scope included only systems that could detect speech and transcribe it in natural and structured fashion simultaneously with the doctor-patient interaction. Studies dealing with ASR for transcription of exam results and for registration of the encounter by the physician afterwards were not included since this constitutes a speech-to-text only solution. That has narrowed significantly our results as the later application is far more developed and widespread than our main focus of interest.

The objective of our research was to identify current developments and challenges in regard to this technology because we have assumed that it could significantly change daily clinical practice by keeping the doctor's hand off from the keyboard and his eyes away from the screen and allow his attention to be fully on the patient in front of him. In that sense, we have aimed to find not only papers that described how the intelligent systems worked and were developed, but also those that reported about their usability, acceptability and accuracy. Finally, data on real-life testing and clinical validation were highly anticipated as evidence of the potential impact of the interventions for improved patient care.

As presented above, few papers were found that properly described applications with the aforementioned characteristics. Only eight publications were selected as most other articles were considered to be about speech-to-text only systems. The programs were in different stages of development and at least one was created for a very specific context (Woo et al., 2021). The intelligent models were presented in more details above, but we can summarize that most consisted in an ASR system with natural language processing capability, a medical lexicon and structured text output. Importantly, documentation should be presented in a structured manner and not as a plain dialogue transcription.

The solutions described have not been clinically validated and many were in its early stages of development. Some have not even presented tests reporting user acceptability or word accuracy (Maas et al., 2021; Klann and Szolovits, 2009; Finley et al., 2018). Two of them were tested a single time by one doctor (Wenceslao and Estuar, 2019; Ni et al., 2014) and another used only recorded audio from patient-doctor dialogues (Khattak et al., 2019). The two most thoroughly tested applications involved real patients in an ophthalmology clinic (Ahamed et al., 2021) or simulations of rescue situations (Woo et al., 2021). We can easily depict from the examples above that clinical data on acceptability, time saved by physicians using these applications, user satisfaction and accuracy in real-world clinical scenarios is absent from the literature.

Future developments can already be foreseen. First, the new technologies should be tested in real-life clinical scenarios and efficiency should be evaluated using larger numbers of patients. Maybe a clinic, a hospital or a health-care system would be suitable environments to test how the application could change everyday work for better or for worse. Second, more work will be required to improve the technical issues that will arise after deployment to real-life interactions. Recognizing the different individuals when more people are talking at the same time, protecting against environmental noise and interpreting medical jargon are difficulties that have appeared in almost all instances so far.

Third, automatic speech recognition and automatic structured documentation could be integrated with other artificial intelligence solutions to improve health care. Machine learning algorithms could be employed to identify red flags, suggest important questions or interventions and to fill paperwork faster. The final goal is more safety to the patient and more time to the clinician expend interacting with patients.

7. Conclusion

Automatic speech recognition may be a valuable tool in the future to facilitate medical register in a faster and more reliable manner. By improving transparency, accuracy and empathy, it could drastically change the way patients and doctors experience a medical visit. Some products that use this technology are already been commercialized, but lacking the scientific evidence of their usability, acceptability and accuracy. Herein we have summarized the latest research on the subject that could be found in the literature. Unfortunately, clinical data on usability and benefits of such applications is almost non-existent. We believe that future work in this area is necessary and needed.

8. Acknowledgment

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel - CAPES (Finance Code 001) and the National Council for Scientific and Technological Development - CNPq (Grant Numbers 309537/2020-7 and 404572/2021-9) for supporting this work.

References

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon 4, e00938. doi:https://doi.org/10.1016/j.heliyon.2018.e00938.

Adler-Milstein, J., Zhao, W., Willard-Grace, R., Knox, M., Grumbach, K., 2020. 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. Journal of the American Medical Informatics Association 27, 531–538. doi:https://doi.org/10.1093/jamia/ocz220.

Ahamed, S., Weiler, G., Boden, K., Januschowski, K., Stennes, M., McCrae, P., Bock, C., Rawein, C., Petris, M., Foth, K., et al., 2021. Deep neural network driven speech classification for relevance detection in automatic medical documentation, in: Public Health and Informatics. IOS Press, pp. 63–67. doi:10.3233/SHTI210121.

Alapetite, A., Andersen, H.B., Hertzum, M., 2009. Acceptance of speech recognition by physicians: A survey of expectations, experiences, and social influence. International journal of human-computer studies 67, 36–49. doi:https://doi.org/10.1016/j.ijhcs.2008.08.004.

Aldosari, B., 2017. Patients' safety in the era of emr/ehr automation. Informatics in Medicine Unlocked 9, 230–233. doi:https://doi.org/10.1016/j.imu.2017.10.001.

Bell, S.K., Delbanco, T., Elmore, J.G., Fitzgerald, P.S., Fossa, A., Harcourt, K., Leveille, S.G., Payne, T.H., Stametz, R.A., Walker, J., DesRoches, C.M., 2020. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. JAMA Network Open 3, e205867–e205867. doi:https://doi.org/10.1001/jamanetworkopen.2020.5867.

Biolchini, J., Mian, P.G., Natali, A.C.C., Travassos, G.H., 2005. Systematic review in software engineering. System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES 679, 45.

Blackley, S.V., Huynh, J., Wang, L., Korach, Z., Zhou, L., 2019. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. Journal of the american medical informatics association 26, 324–338. doi:https://doi.org/10.1093/jamia/ocy179.

Brooke, J., et al., 1996. Sus – a quick and dirty usability scale. Usability evaluation in industry 189, 4–7.

van Buchem, M.M., Boosman, H., Bauer, M.P., Kant, I.M., Cammel, S.A., Steyerberg, E.W., 2021. The digital scribe in clinical practice: a scoping review and research agenda. NPJ digital medicine 4, 1–8. doi:https://doi.org/10.1038/s41746-021-00432-5.

Callaway, E.C., Sweet, C.F., Siegel, E., Reiser, J.M., Beall, D.P., 2002. Speech recognition interface to a hospital information system using a self-designed visual basic program: initial experience. Journal of digital imaging 15, 43–53. doi:https://doi.org/10.1007/BF03191902.

Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K., 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. PLoS computational biology 9, e1002854.

Eisenberg, M.J., 1995. Accuracy and predictive values in clinical decision-making. Cleveland Clinic journal of medicine 62, 311–316. doi:https://doi.org/10.3949/ccjm.62.5.311.

Finley, G.P., Edwards, E., Robinson, A., Sadoughi, N., Fone, J., Miller, M., Suendermann-Oeft, D., Brenndoerfer, M., Axtmann, N., 2018. An automated assistant for medical scribes., in: INTERSPEECH, pp. 3212–3213. doi:10.21437/Interspeech.2018.

Ghatnekar, S., Faletsky, A., Nambudiri, V.E., 2021. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. Health and Technology , 1–7doi:https://doi.org/10.1007/s12553-021-00568-0.

Groopman, J., 2008. How doctors think. Houghton Mifflin Harcourt.

Gupta, D., Bansal, P., Choudhary, K., 2018. The state of the art of feature extraction techniques in speech recognition. Speech and language processing for human-machine communications , 195–207doi:https://doi.org/10.1007/978-981-10-6626-9_22.

Hamet, P., Tremblay, J., 2017. Artificial intelligence in medicine. Metabolism 69, S36–S40. doi:https://doi.org/10.1016/j.metabol.2017.01.011.

Hammond, K.W., Helbig, S.T., Benson, C.C., Brathwaite-Sketoe, B.M., 2003. Are electronic medical records trustworthy? observations on copying, pasting and duplication, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 269.

Heart, T., Ben-Assuli, O., Shabtai, I., 2017. A review of phr, emr and ehr integration: A more personalized healthcare and public health policy. Health Policy and Technology 6, 20–25. doi:https://doi.org/10.1016/j.hlpt.2016.08.002.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology 2. doi:https://doi.org/10.1136/svn-2017-000101.

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., Dawson, L., 2014. A systematic review of speech recognition technology in health care. BMC medical informatics and decision making 14, 1–14. doi:https://doi.org/10.1186/1472-6947-14-94.

Khattak, F.K., Jeblee, S., Crampton, N., Mamdani, M., Rudzicz, F., 2019. Autoscribe: extracting clinically pertinent information from patient-clinician dialogues, in: MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press BV, pp. 1512–1513. doi:https://doi.org/10.3233/SHTI190510.

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

Klann, J.G., Szolovits, P., 2009. An intelligent listening framework for capturing encounter notes from a doctor-patient dialog. BMC medical informatics and decision making 9, 1–10. doi:https://doi.org/10.1186/1472-6947-9-S1-S3.

Kroth, P.J., Morioka-Douglas, N., Veres, S., Babbott, S., Poplau, S., Qeadan, F., Parshall, C., Corrigan, K., Linzer, M., 2019. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Network Open 2, e199609–e199609. doi:https://doi.org/10.1001/jamanetworkopen.2019.9609.

Latif, S., Qadir, J., Qayyum, A., Usama, M., Younis, S., 2021. Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering 14, 342–356. doi:https://doi.org/10.1109/RBME.2020.3006860.

Lipkin, M., 1987. The medical interview and related skills .

Maas, L., Kisjes, A., Hashemi, I., Heijmans, F., Dalpiaz, F., Dulmen, S.V., Brinkkemper, S., 2021. Automated medical reporting: From multimodal inputs to medical reports through knowledge graphs, in: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: HEALTHINF,, INSTICC. SciTePress. pp. 509–514. doi:10.5220/0010261605090514.

Margalit, R.S., Roter, D., Dunevant, M.A., Larson, S., Reis, S., 2006. Electronic medical record use and physician–patient communication: An observational study of israeli primary care encounters. Patient Education and Counseling 61, 134–141. doi:https://doi.org/10.1016/j.pec.2005.03.004.

Matar Boumosleh, J., Jaalouk, D., 2017. Depression, anxiety, and smartphone addiction in university students- a cross sectional study. PLOS ONE 12, 1–14. doi:https://doi.org/10.1371/journal.pone.0182239.

Ni, G., Shi, W., Mahajan, P., 2014. Appurtenant: enhancing completeness and efficiency of bidirectional patient-physician communication using automatic speech recognition, in: Proceedings of the 2014 workshop on Mobile augmented reality and robotic technology-based systems, pp. 35–40. doi:10.1145/2609829.2609830.

Orient, J.M., Sapira, J.D., 2012. Sapira's Art & Science of Bedside Diagnosis. LWW.

Otter, D.W., Medina, J.R., Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems 32, 604–624. doi:https://doi.org/10.1109/TNNLS.2020.2979670.

Panch, T., Szolovits, P., Atun, R., 2018. Artificial intelligence, machine learning and health systems. Journal of global health 8. doi:https://doi.org/10.7189/jogh.08.020303.

Paranjape, K., Schinkel, M., Panday, R.N., Car, J., Nanayakkara, P., 2019. Introducing artificial intelligence training in medical education. JMIR medical education 5, e16048. doi:https://doi.org/10.2196/16048.

Payne, T.H., Alonso, W.D., Markiel, J.A., Lybarger, K., Lordon, R., Yetisgen, M., Zech, J.M., White, A.A., 2018. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. Jamia Open 1, 218–226. doi:10.1093/jamiaopen/ooy036.

Pearce, C., Trumble, S., Arnold, M., Dwan, K., Phillips, C., 2008. Computers in the new consultation: within the first minute. Family Practice 25, 202–208. doi:https://doi.org/10.1093/fampra/cmn018.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., Duffy, S., et al., 2006. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version 1, b92.

Prevedello, L.M., Ledbetter, S., Farkas, C., Khorasani, R., 2014. Implementation of speech recognition in a community-based radiology practice: effect on report turnaround times. Journal of the American College of Radiology 11, 402–406. doi:https://doi.org/10.1016/j.jacr.2013.07.008.

Quiroz, J.C., Laranjo, L., Kocaballi, A.B., Berkovsky, S., Rezazadegan, D., Coiera, E., 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ digital medicine 2, 1–6. doi:https://doi.org/10.1038/s41746-019-0190-1.

Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. New England Journal of Medicine 380, 1347–1358. doi:https://doi.org/10.1056/NEJMra1814259.

Roehrs, A., Da Costa, C.A., da Rosa Righi, R., De Oliveira, K.S.F., 2017. Personal health records: a systematic literature review. Journal of medical Internet research 19, e13. doi:https://doi.org/10.2196/jmir.5876.

Sackett, D.L., 1992. A primer on the precision and accuracy of the clinical examination. Jama 267, 2638–2644. doi:https://doi.org/10.1001/jama.1992.03480190080037.

Schnabel, T.G., 1983. Is medicine still an art?

Scott, D., Purves, I.N., 1996. Triadic relationship between doctor, computer and patient. Interacting with Computers 8, 347–363. doi:https://doi.org/10.1016/S0953-5438(97)83778-2.

Singh, M., Pal, T.R., 2011. Voice recognition technology implementation in surgical pathology: advantages and limitations. Archives of pathology & laboratory medicine 135, 1476–1481. doi:https://doi.org/10.5858/arpa.2010-0714-OA.

Sohn, S.Y., Rees, P., Wildridge, B., Kalk, N.J., Carter, B., 2019. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and grade of the evidence. BMC Psychiatry 19, 356. doi:https://doi.org/10.1186/s12888-019-2350-x.

Stevenson, I., 1971. The diagnostic interview. New York: Medical Department, Harper & Row.

Topol, E., 2019. Deep medicine: how artificial intelligence can make healthcare human again. Basic Books USA.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J., 2004. Sphinx-4: A flexible open source framework for speech recognition.

Weinstein, M.C., Fineberg, H.V., Elstein, A.S., Frazier, H.S., Neuhauser, D., Neutra, R.R., McNeil, B.J., 1980. Clinical decision analysis. Saunders Philadelphia.

Wenceslao, S.J.M.C., Estuar, M.R.J.E., 2019. Using ctakes to build a simple speech transcriber plugin for an emr, in: Proceedings of the Third International Conference on Medical and Health Informatics 2019, Association for Computing Machinery, New York, NY, USA. pp. 78–86. doi:https://doi.org/10.1145/3340037.3340044.

Woo, M., Mishra, P., Lin, J., Kar, S., Deas, N., Linduff, C., Niu, S., Yang, Y., McClendon, J., Smith, D.H., et al., 2021. Complete and resilient documentation for operational medical environments leveraging mobile hands-free technology in a systems approach: Experimental study. JMIR mHealth and uHealth 9, e32301. doi:https://doi.org/10.2196/32301.

Zick, R.G., Olsen, J., 2001. Voice recognition software versus a traditional transcription service for physician charting in the ed. The American journal of emergency medicine 19, 295–298. doi:https://doi.org/10.1053/ajem.2001.24487.

Ziman, K., Heusser, A.C., Fitzpatrick, P.C., Field, C.E., Manning, J.R., 2018. Is automatic speech-to-text transcription ready for use in psychological experiments? Behavior research methods 50, 2597–2605. doi:https://doi.org/10.3758/s13428-018-1037-4.

Highlights

- We survey intelligent models for the automatic documentation of health interactions
- The methodology consists of a systematic literature review of the last 10 years
- Solutions based on neural networks were the most used
- The main challenges are the complexity of vocabulary/ relevant parts identification
- Legal, ethical, and moral challenges are key factors for healthcare systems

Fig. 3

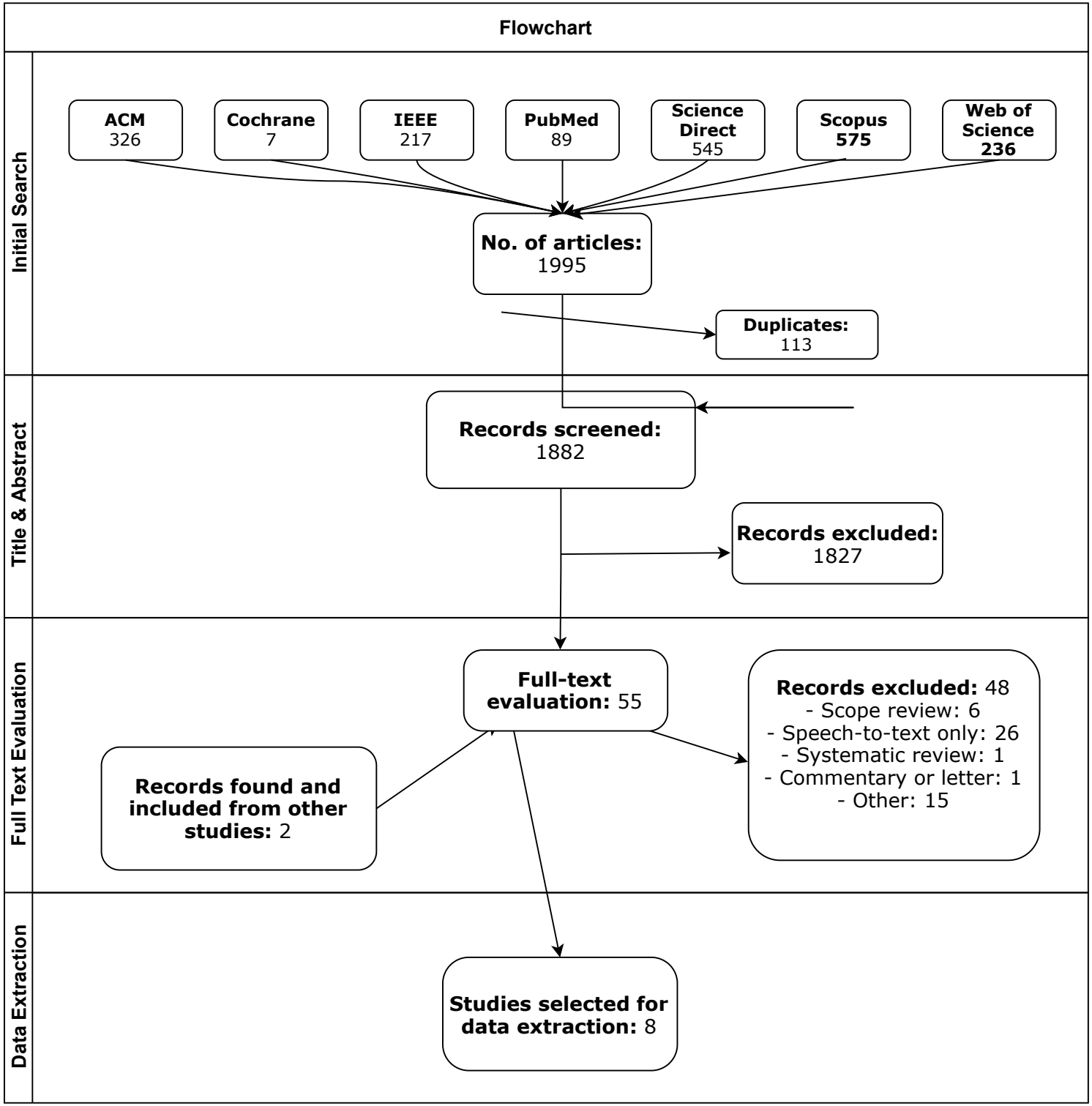
Flowchart
Initial Search
ACM
326
Cochrane
7
IEEE
217
PubMed
89
Science Direct
545
Scopus
575
Web of Science
236
No. of articles:
1995
Duplicates:
113
Title & Abstract
Records screened:
1882
Records excluded:
1827
Full Text Evaluation
Full-text evaluation: 55
Records excluded: 48
- Scope review: 6
- Speech-to-text only: 26
- Systematic review: 1
- Commentary or letter: 1
- Other: 15
Records found and included from other studies: 2
Data Extraction
Studies selected for data extraction: 8

Fig. 2

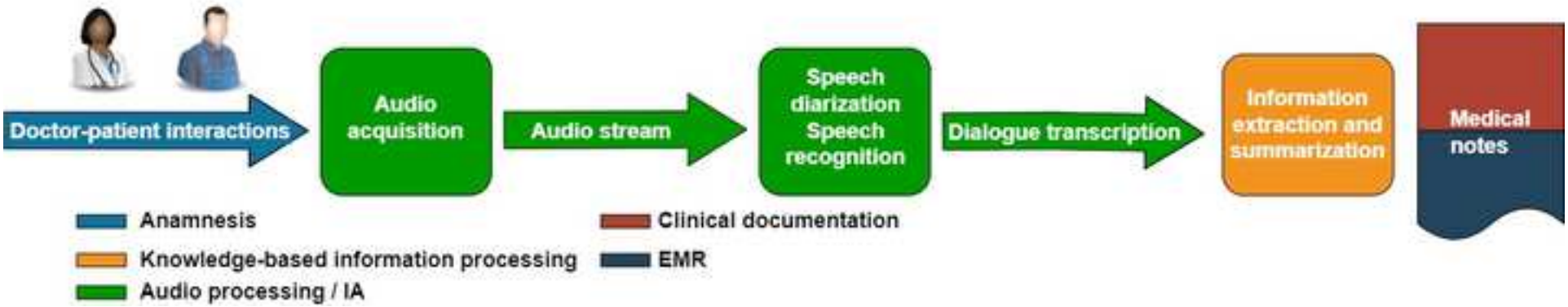
Doctor-patient interactions
Audio acquisition
Audio stream
Speech diarization Speech recognition
Dialogue transcription
Information extraction and summarization
Medical notes
Anamnesis
Clinical documentation
Knowledge-based information processing
EMR
Audio processing / IA

Fig. 1

- Structured and dynamic process
- Source of clinical information
- Use of a theoretical and practical domains

- Based on clinical knowledge
- Information summarization
- Clinical reasoning for diagnosis and treatment

- Structured registration
- Details medical information
- Communicate professional impressions and actions

- Digital version of paper chart
- Primary source of patient clinical information
- Where health care providers share information

Automatic documentation of professional health interactions: a systematic review

Frederico Soares Falcetta, MD, M.Sc. in Pathology
Software Innovation Laboratory - SOFTWARELAB
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: fredfalcetta@gmail.com
ORCID: 0000-0002-6457-4164

Fernando Kude de Almeida, MD, M.Sc. in Pathology
Hospital Fêmina - Porto Alegre, Brazil
E-mail: fernandokude@gmail.com
ORCID: 0000-0003-2847-0664

Janaína Conceição Sutil Lemos, M.Sc. in Computer Science
Escola Politécnica
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: csutil@unisinos.br
ORCID: 0000-0003-1105-1260

José Roberto Goldim, Ph.D. in Medicine, Professor
Bioethics Division, Hospital de Clínicas de Porto Alegre/Brazil, Porto Alegre, Brazil.
E-mail: jgoldim@hcpa.edu.br
ORCID: 0000-0003-2127-6594

[*]Cristiano André da Costa, Ph.D. in Computer Science, Professor
Software Innovation Laboratory - SOFTWARELAB
Universidade do Vale do Rio dos Sinos - Unisinos - São Leopoldo, Brazil
E-mail: cac@unisinos.br
ORCID: 0000-0003-3859-6199

[*]Corresponding author

Software Innovation Laboratory - SOFTWARELAB

Programa de Pós-Graduação em Computação Aplicada

Universidade do Vale do Rio dos Sinos

Av. Unisinos 950 93022-000 São Leopoldo RS, Brazil
Phone: +55 51 35908161 Fax: +55 51 35908162

São Leopoldo, Brazil, April 12th, 2022.

CONFLICT OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with the article "Automatic documentation of professional health interactions: a systematic review" sent to possible publication in Expert Systems with Applications and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Prof. Cristiano André da Costa, PhD
(Corresponding Author, in the name of the authors)

Address: Programa de Pós-graduação em Computação Aplicada
Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos, 950
São Leopoldo, RS, Brazil, 93022-750
Phone/ FAX: +55(51) 3590-8161 / +55(51) 35908162

E-mail: cac@unisinos.br / caccac@gmail.com